Advancing Biogeographical Ancestry Predictions Through Machine Learning

Carola Sophia Heinzel[1], Lennart Purucker[2], Frank Hutter[4,3,2], Peter Pfaffelhuber[1]
(1) Department of Mathematical Stochastics, University of Freiburg
(2) Machine Learning Lab, University of Freiburg
(3) PriorLabs (4) ELLIS Institute Tübingen

January 31, 2025

ABSTRACT: Tools like Snipper *or the Admixture Model count as state-of-the-art methods in forensic science for biogeographical ancestry. However, they have not been systematically compared to classifiers widely used in other disciplines. Noting that genetic data have a tabular form, this study addresses this gap by benchmarking forensic classifiers against* TabPFN*, a cutting-edge, general-purpose machine learning classifier for tabular data. The comparison evaluates performance using metrics such as accuracy—the proportion of correct classifications—and ROC AUC. We examine classification tasks for individuals at both the intracontinental and continental levels, based on a published dataset for training and testing. Our results reveal significant performance differences between methods, with* TabPFN *consistently achieving the highest accuracy and ROC AUC across the datasets. Based on these findings, we recommend adopting* TabPFN *for population classification in forensic science.*

1 Introduction

Classification of individuals into ancestral populations, i.e., inferring biogeographical ancestry (BGA) from DNA traces, is a fundamental task in forensic science (Wen et al., 2023). This process is particularly useful for identifying disaster victims (Phillips et al., 2009) and in broader forensic investigations (Phillips et al., 2007b). Additional applications include detecting population stratification (Dang et al., 2005; Min et al., 2000; Gan et al., 2003). Researchers commonly classify individuals into several continental populations such as Africa, Europe, East Asia, South Asia, Admixed Americans, and Oceania (Cheung et al., 2017; Wang et al., 2007), or on an intracontinental level.

The statistical task of classification can be broken down into two steps. First, genetic markers have to be chosen which are able to distinguish between populations, i.e. with frequency differentials between populations. Second, these Ancestry Informative Markers (or AIMs) are

used as features in a classification task. Here, various algorithms are used, leveraging genetic distance, regression, or Bayesian inference (Cheung et al., 2017). Tools like the naive Bayes classifier Snipper (Phillips et al., 2014) have been used alongside multinomial logistic regression (McNevin et al., 2013) and the Genetic Distance Algorithm (Frudakis et al., 2003). Furthermore, the Admixture Model, as implemented in tools like Structure (Pritchard et al., 2000) or Admixture (Alexander et al., 2009), is widely used for classification (Cheung et al., 2017). More recently, Alladio et al. (2022) have applied some machine learning methods, i.e., XGBoost and Partial Least Squares-Discriminant Analysis (PLS-DA).

Several challenges must be taken into account for BGA. First, classification becomes increasingly challenging for more genetically similar populations, which are frequently found within continents. Second, admixed individuals, such as Americans, present difficulties in classification (Lao et al., 2010; Llamas et al., 2016). Third, marker selection significantly affects classification quality (Cheung et al., 2017), with ideal markers showing large differences in allele frequencies between populations (Rosenberg et al., 2003). Fourth, selecting populations as classes must lead to results which are both, reliable and helpful in further investigations.

The choice of Ancestry Informative Markers (AIMs) is frequently performed by finding alleles with large frequency differentials between populations using expert knowledge (Kidd et al., 2014; Phillips et al., 2014; Ruiz-Ramírez et al., 2023). However, feature selection for classification is also a well-known task in machine learning, which has been applied to the choice of AIMs (Wathen et al., 2019; Pfaffelhuber et al., 2020b). The goal is to find a set which leads to a reduced sequencing effort, focusing on 50–200 informative markers (relative to the classes under study) instead of thousands (Ko et al., 2023). However, Heinzel et al. (2024) showed that marker quality criteria extend beyond allele frequency differences. Additionally, forensic contexts often face challenges with missing data (Cheung et al., 2017).

Cheung et al. (Cheung et al., 2017) compared several classification approaches, including Snipper (Phillips et al., 2014), multinomial logistic regression (McNevin et al., 2013), and the Genetic Distance Algorithm (Frudakis et al., 2003). They also analyzed the Admixture Model, concluding that "*STRUCTURE was the most accurate classifier for both complete and partial genotypes in non-admixed individuals across four reference populations, including populations with suspected admixture.*" This study is complemented by a comparison to XGboost and PLS-DA in (Alladio et al., 2022), who conclude that such machine learning tools have a higher accuracy when compared to the above methods. We complement this study by using TabPFN, a new foundational model which was shown to outperform existing models for tabular data (Hollmann et al., 2025).

Ideally, a classifier predicts the population of every individual with 100% ancestry of one population correctly with prediction probability 1. Additionally, according to Cheung et al. (2017), for admixed individuals the prediction probability should be equal to the ancestry proportions of the genome. Several measures of non-optimal classifiers, such as the misclassification rate, the logloss (sum of logs of misclassification probabilities), accuracy (fraction of correctly classified samples), the ROC AUC (the area under a curve of false positives against true positives) are among the most common statistics.

Machine learning has already seen applications in forensic science, including haplogroup prediction from Y-STR data (Swaminathan et al., 2015), estimation of the number of contributors to a trace (Bouakaze et al., 2020), and prediction of the age or post-mortem interval (Barash et al., 2024). While machine learning methods have been employed to predict physical appearance from DNA (Katsara et al., 2021) as well as BGA (Wathen et al., 2019), the application of such methods have been criticized for their lack of fit to the specific needs in the forensic field (Barash et al., 2024). Our work aims to address this gap by applying state-of-the-art machine learning techniques to the BGA classification problem.

2 Material and Methods

2.1 Choice of the Data and the Marker Set

Throughout our study, we adopt the dataset from Ruiz-Ramírez et al. (2023) (their Supplementary Table S1A), which uses the markers from the VISAGE Enhanced tool (Xavier et al., 2020). This set of markers has 104 autosomal markers, 29 of which are multiallelic. This dataset consists of 2504 individuals from the 1000 genomes project (Consortium et al., 2015), 929 HGDP-CEPH samples (Phillips et al., 2013; Santos et al., 2016), 137 samples from the Middle East (Almarri et al., 2021), 130 samples from the SGDP (Mallick et al., 2016), and 402 samples from the Estonian Biocentre human genome diversity panel (Pagani et al., 2016).

We considered two cases:

(1) Continental level combining *'AFRICAN'*,*'EAST AFRICAN'* and *'ADMIXED AFRICAN'* to Africa (AFR), combining *'MIDDLE EAST'* and *'(SANGER) MIDDLE EAST'* to Middle East (ME), combining *'EUROPEAN'* and *'ROMA'* to Europe (EUR), East Asia (EAS), South-East Asia (SAS), Oceania (OCE), *'CENTRAL ASIAN'* (CAS), *'NORTH AFRICAN'* (NAF) and combining *'ADMIXED AMERICAN'* and *'AMERICAN'* to Admixed America (AMR). This makes a total of 9 classes.

(2) Intra-European level, only considering the EUR classes from above with more than 20 individuals, using *(CEPH) with N & W European ancestry'* (CEU), *'Finnish in Finland'* (FIN), *'British in England and Scotland'* (GBR), *'Iberian population in Spain'* (IBS), combining *'Toscani in Italia' 'Italy - Sardinian'* (ITA), *'Russia - Russian'* (RUS), combining *'France - French Basque'* and *'France - French'* (FRA), and *'Turkey'* (TUR) as classes. These are a total of 8 classes.

The number of individuals is shown for both cases in the Figure 1, where we see an extremely different number of individuals in both cases.

2.2 Classification

For classification, we compare the following methods: (i) Snipper, which is a version of a naive Bayes classifier (Phillips et al., 2007a); (ii) the Admixture Model (AM) (in its supervised

⟨fig:case_a⟩ (A) (B)

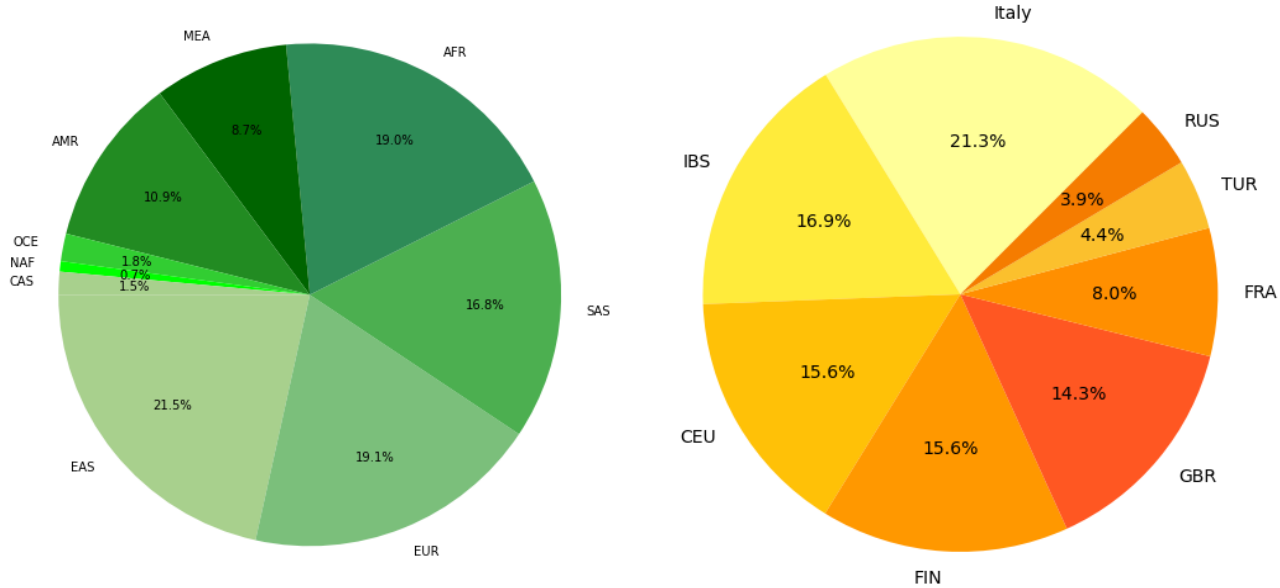

Figure 1: (A) Overview of the number of individuals per population in case (1), i.e. for continental classification. In this case, there are a total of 4342 individuals. (B) Overview of the number of individuals per population for case (2), i.e. intra-European classification. In this case, there are a total of 635 individuals.

⟨fig:case_a⟩

setting), which assumes that every allele in individual i has a chance of q_{ik} to come from population k (Pritchard et al., 2000; Alexander et al., 2009), and individual i is classified into population k if q_{ik} is maximal; (iii) PLS-DA which has already been used by (Alladio et al., 2022) in the forensic context; and (iv) TabPFN, a novel foundational model for tabular data (such as genetic data) based on a neural network (Hollmann et al., 2025). Note that the model-based methods (i) and (ii) treat markers as independent, while the machine-learning (black-box) methods (iii) and (iv) incorporate joint segregation of markers.

Other classifiers, such as gradient-boosted decision trees (Chen and Guestrin; Prokhorenkova et al.; Ke et al.), or logistic regression (Cheung et al., 2019), would as well be applicable to genetic (i.e., tabular) data. However, Hollmann et al. (2025) have shown that TabPFN dominantly outperforms other classifiers, including those mentioned, for small tabular datasets.

Neural networks play a central role in machine learning today. While convolutional neural networks have become the foundation for models for image (and video) data, the transformer neural network architecture is the foundation for large language models (such as ChatGPT). With TabPFN, a foundation model for tabular data was introduced, making it possible to apply transformer neural networks to any classification problem on tables instead of images or text. We stress that TabPFN is a pre-trained neural network, and the weights (which are independent of the application context) are open and free to use[1].TabPFN excels in handling small to medium-sized datasets with up to 10,000 samples and 500 features.

[1] https://github.com/PriorLabs/tabpfn

TabPFN has a higher usage cost than traditional classifiers because it needs a graphics processing unit (GPU) to predict efficiently. However, TabPFN is a relatively small model (as compared to large language models), since it contains millions of parameters rather than billions (or even trillions). As a result, it can run efficiently on consumer-grade hardware, such as desktop GPUs or integrated GPUs (e.g., Apple's M-series chips), with about 8GB of VRAM. Despite the usage cost, we select TabPFN for our comparison because it requires substantially less training time if a GPU is available (Hollmann et al., 2025).

We evaluate each classifier by training the model on a subset of the data and evaluating its prediction on a holdout set. This experiment is repeated 50 times by using 10-repeated 5-fold cross-validation with stratification from scikit-learn (Pedregosa et al., 2011). We score predictions using accuracy, log loss and ROC AUC[2], following best practices from machine learning (Provost et al., 1998; McElfresh et al.; Gijsbers et al., 2024; Vovk, 2015). ROC AUC measures the ability to distinguish between classes and 1, indicating perfect discrimination, and 0.5, representing random guessing. Analogously to previous evaluation metrics in the forensics (Pfaffelhuber et al., 2020a; Resutik et al., 2023), we also count the number of true positive and false negative samples to compute a confusion matrix. More precisely, for every one of our 50 repetitions, we sum up the number of individuals from class i, e.g., Europe, that are classified into population j, e.g., Africa. The results are depicted as a confusion matrix.

2.3 Code availability

All code for our analysis can be downloaded from GitHub. We include mostly Python scripts for selecting individuals from the dataset, and a locally running web-interface for easier use of classification. We also set up this interface online.

3 Results

First, we compare the evaluation metrics ROC AUC, Accuracy and Log Loss between the different classifiers and between the continental and intracontinental classification. Second, we compare the corresponding confusion matrices of the classification methods to each other. We provide results for all combinations of classifier and continental and inter-European classification.

3.1 Evaluation Metrics

The ROC AUC (between 0.5 and 1, higher is better), Accuracy (between 0 and 1, higher is better) and the Log loss (greater than 0, lower is better) of the classification in case (A) and (B) can be found in Figure 2.

For all three evaluation metrics TabPFN is the best classification method in both, continental and inter-European classification. While performing worse than TabPFN, PLS-DA

[2]We employ scikit-learn's one-vs-rest method to extend ROC AUC to multiclass.

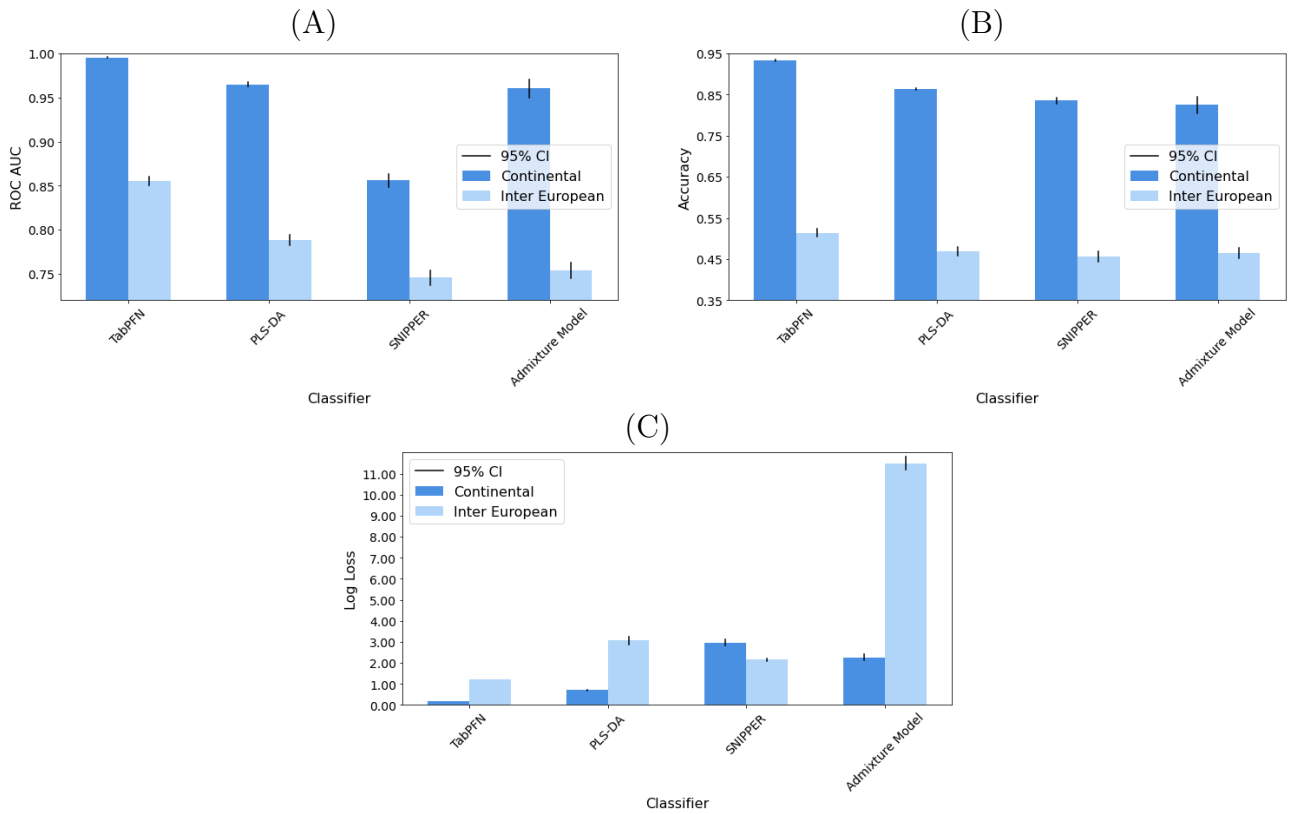

Figure 2: Evaluation metrics for continental (1) and inter-European (2) classification. The ⟨metric⟩ black line represents the 95% confidence interval. (A) ROC AUC; (B) Accuracy; (C) Log Loss.

is better than Snipper and the Admixture Model in most cases. ROC AUC for Snipper is worse than for Admixture Model while the accuracy and the log loss of Snipper are better than for the Admixture Model. Altogether, all four considered classification methods work well for continental classification. Unsurprisingly, performance drops significantly for inter-European classification. Even TabPFN only classifies 53% of samples correctly, which can be explained by the relatively large number of classes (8) and the high similiarity between the classes.

3.2 Confusion Matrices

For more precise results which distinctions work well and which don't, we display confusion matrices for all eight cases. (Cases (1) and (2), four classifiers.) The results are shown in Figures 3 and 4.

For continental classification, we see that the classification works very well for all populations except for ME and NAF in all four methods. In 81% of the considered cases, TabPFN classifies ME individuals correctly, which is only outperformed by Snipper (88%). However, Snipper misclassifies the majority of NAF into ME (only 17% being correct), whereas TabPFN achieves 88% also in NAF. PLS-DA and the Admxiture Model are similar to Snipper in the confusion between ME and NAF. Another interesting population is Central Asian. (Note that

this data is from EGDP (Pagani et al., 2016) and only contains 64 samples.) Here, TabPFN only achieves 61%, whereas methods which treat markers independently (Snipper, Admixture Model) achieve 100%.

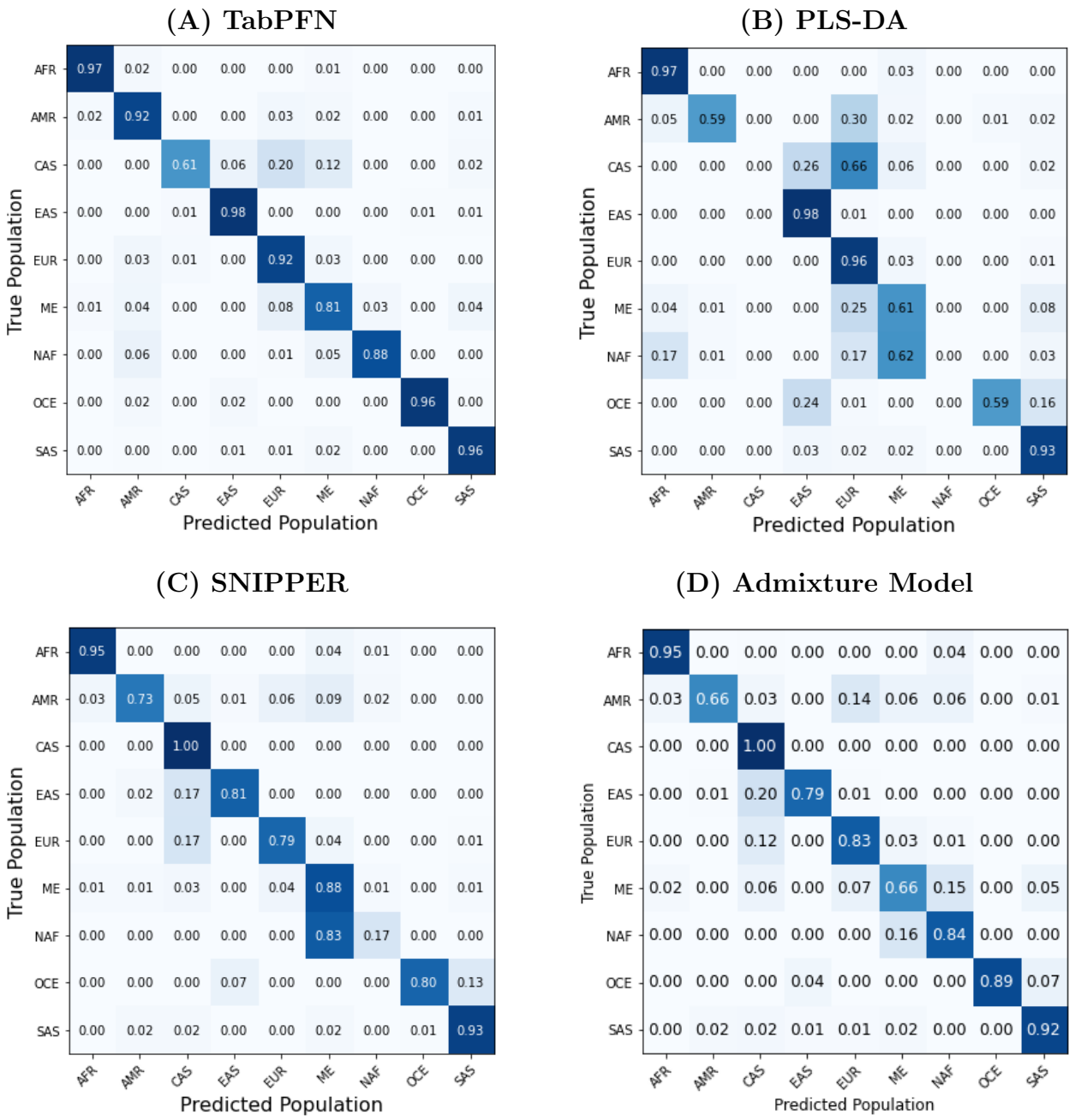

Figure 3: Confusion matrices for continental classification (case 1). All rows sum to 1 up to a rounding error. **(A)** TabPFN; **(B)** PLS-DA; **(C)** SNIPPER; **(D)** Admixture Model.

fusion_cont⟩

Moreover, we show the confusion matrix for intra-European classification (case 2). With TabPFN, we see that people from CEU are often (in 25% of the cases considered) classified into GBR. Furthermore, people from GBR are often classified as CEU (in 48% of the cases

considered). The number of miss-classifications for French people is in general very high and the number of miss-classifications of RUS into FIN is also large.

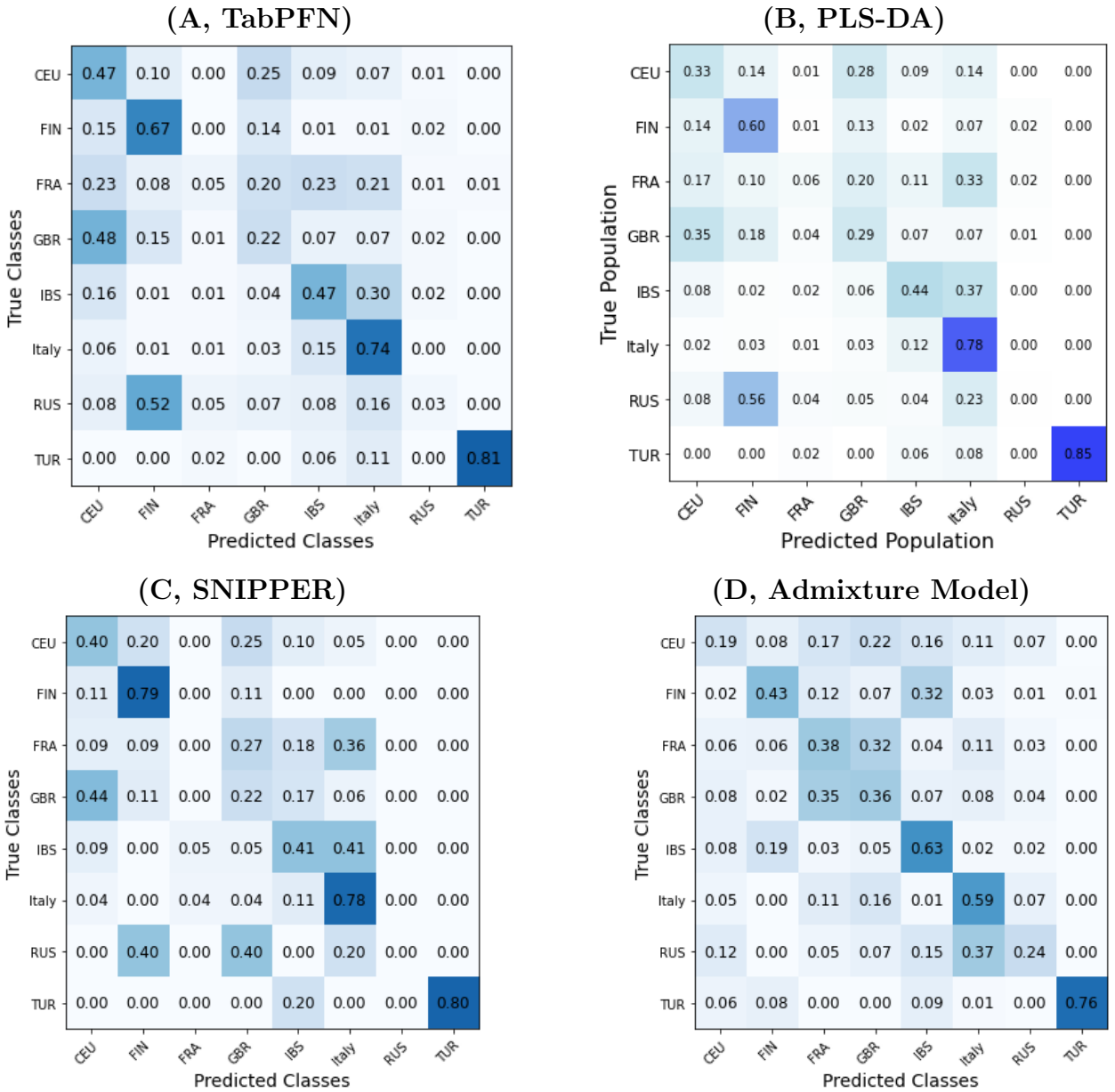

Figure 4: Confusion matrices for intra-European classification (case 2). All rows sum to 1 up
nfusion_EUR⟩ to a rounding error. (A) TabPFN; (B) PLS-DA; (C) Snipper; (D) Admixture Model

In comparison to TabPFN, the percentage of correct classifications of people from some populations is even higher for Snipper or for the Admixture Model. An example for this are the individuals from Finland with Snipper and the ones from GBR with the Admixture Model.

4 Discussion

In our study, we test four classifiers on well-established (training and test) data, which was evaluated in the VISAGE project (Xavier et al., 2022). We only use classifiers which determine their own uncertainty by reporting probabilities for class membership. Our results show that TabPFN produces the highest average ROC AUC, the highest average accuracy, and the best average log loss for our data. Especially when the populations are exclusively from Europe, TabPFN strongly outperforms the two state-of-the-art methods and PLS-DA.

Hence, we advocate the use of TabPFN to classify individuals into populations. Therefore, we provide an offline interface with some example data. However, the user is free to use the method on any other data. It is worth noting that on an intracontinental level, TabPFN outperforms other methods, but is still not good enough for a highly accurate classification. To evaluate whether a classification into a specific choice of populations achieves high accuracy, researchers should evaluate this on a data set with known ancestry. In addition, if the training data has only a few samples for some classes, accuracy drops due to the increased sampling variance.

Furthermore, at the time of writing, the latest version of TabPFN is recommended for use on data with up to 500 markers, 10000 individuals, and 10 classes. However, we believe that these are only temporary limitations. The machine learning community already produced many extensions of TabPFN [3] and is exploring means to scale tabular foundation models like TabPFN to support more markers, individuals, and classes (Ma et al., 2024a; Zeng et al., 2024; Feuer et al., 2024; Ma et al., 2024b).

Improving the performance of classification in BGA suggests that similar tasks can be improved as well. In particular, researchers can apply TabPFN to external visible characteristics, where eye color (Liu et al., 2009; Walsh et al., 2011) or the hair color prediction (Walsh et al., 2013) are considered. Moreover, we used TabPFN without optimizing (i) the input for the markers used or (ii) the weights of the neural network for the specific task. These two possibilities might even further increase the accuracy, although they come with a large computational cost.

5 Funding

We acknowledge funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under SFB 1597 (SmallData), grant number 499552394. This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under grant number 417962828. Frank Hutter acknowledges the financial support of the Hector Foundation.

[3] See for example `https://github.com/PriorLabs/tabpfn-extensions`.

6 Conflicts of interest

Frank Hutter cofounded the tabular foundation company PriorLabs that open-sourced TabPFN and is working on better models. The authors declare that there are no other conflicts of interest.

References

09fast David H Alexander, John Novembre, and Kenneth Lange. Fast model-based estimation of ancestry in unrelated individuals. *Genome research*, 19(9):1655–1664, 2009.

ariate Eugenio Alladio, Brando Poggiali, Giulia Cosenza, and Elena Pilli. Multivariate statistical approach and machine learning for the evaluation of biogeographical ancestry inference in the forensic field. *Scientific Reports*, 12(1):8974, 2022.

ri2021 Mohamed A Almarri, Marc Haber, Reem A Lootah, Pille Hallast, Saeed Al Turki, Hilary C Martin, Yali Xue, and Chris Tyler-Smith. The genomic history of the middle east. *Cell*, 184 (18):4612–4625, 2021.

achine Mark Barash, Dennis McNevin, Vladimir Fedorenko, and Pavel Giverts. Machine learning applications in forensic dna profiling: A critical review. *Forensic Science International: Genetics*, 69:102994, 2024.

icting Caroline Bouakaze, Franklin Delehelle, Nancy Saenz-Oyhéréguy, Andreia Moreira, Stéphanie Schiavinato, Myriam Croze, Solène Delon, Cesar Fortes-Lima, Morgane Gibert, Louis Bujan, et al. Predicting haplogroups using a versatile machine learning program (predymale) on a new mutationally balanced 32 y-str multiplex (combyplex): Unlocking the full potential of the human str mutation rate spectrum to estimate forensic parameters. *Forensic Science International: Genetics*, 48:102342, 2020.

xdd16a T. Chen and C. Guestrin. XGBoost: A scalable tree boosting system. pages 785–794.

iction Elaine YY Cheung, Michelle Elizabeth Gahan, and Dennis McNevin. Prediction of biogeographical ancestry from genotype: a comparison of classifiers. *International journal of legal medicine*, 131:901–912, 2017.

ng2019 Elaine YY Cheung, Christopher Phillips, Mayra Eduardoff, Maria Victoria Lareu, and Dennis McNevin. Performance of ancestry-informative snp and microhaplotype markers. *Forensic Science International: Genetics*, 43:102141, 2019.

global 1000 Genomes Project Consortium et al. A global reference for human genetic variation. *Nature*, 526(7571):68, 2015.

luence Mai-Trang N Dang, Julie Hambleton, and Steven R Kayser. The influence of ethnicity on warfarin dosage requirement. *Annals of Pharmacotherapy*, 39(6):1008–1012, 2005.

tables Benjamin Feuer, Robin Tibor Schirrmeister, Valeriia Cherepanova, Chinmay Hegde, Frank Hutter, Micah Goldblum, Niv Cohen, and Colin White. Tunetables: Context optimization for scalable prior-data fitted networks. *arXiv preprint arXiv:2402.11137*, 2024.

sifier Tony Frudakis, K Venkateswarlu, M Thomas, Zach Gaskin, Siva Ginjupalli, Sitarama Gunturi, Viswanathan Ponnuswamy, Sivamiani Natarajan, and Ponnuswamy Kolathupalayam Nachimuthu. A classifier for the snp-based inference of ancestry. *Journal of forensic sciences*, 48(4):771–782, 2003.

racial Gin Gin Gan, Alan Teh, Kim Yen Goh, Heng Thay Chong, and Kang Wah Pang. Racial background is a determinant factor in the maintenance dosage of warfarin. *International journal of hematology*, 78:84–86, 2003.

mlr24a P. Gijsbers, M. Bueno, S. Coors, E. LeDell, S. Poirier, J. Thomas, B. Bischl, and J. Vanschoren. Amlb: an automl benchmark. 25(101):1–65, 2024.

ealing Carola Sophia Heinzel, Franz Baumdicker, and Peter Pfafffelhuber. Revealing the range of maximum likelihood estimates in the admixture model. *bioRxiv*, pages 2024–10, 2024.

curate Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326, 2025.

uation Maria-Alexandra Katsara, Wojciech Branicki, Susan Walsh, Manfred Kayser, Michael Nothnagel, VISAGE Consortium, et al. Evaluation of supervised machine-learning methods for predicting appearance traits from dna. *Forensic Science International: Genetics*, 53:102507, 2021.

ips17a G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu. Lightgbm: A highly efficient gradient boosting decision tree.

dd2014 Kenneth K Kidd, William C Speed, Andrew J Pakstis, Manohar R Furtado, Rixun Fang, Abeer Madbouly, Martin Maiers, Mridu Middha, Françoise R Friedlaender, and Judith R Kidd. Progress toward an efficient panel of snps for ancestry inference. *Forensic Science International: Genetics*, 10:23–32, 2014.

rvised Seyoon Ko, Benjamin B Chu, Daniel Peterson, Chidera Okenwa, Jeanette C Papp, David H Alexander, Eric M Sobel, Hua Zhou, and Kenneth L Lange. Unsupervised discovery of ancestry-informative markers and genetic admixture proportions in biobank-scale datasets. *The American Journal of Human Genetics*, 110(2):314–325, 2023.

uating Oscar Lao, Peter M Vallone, Michael D Coble, Toni M Diegoli, Mannis van Oven, Kristiaan J van Der Gaag, Jeroen Pijpe, Peter de Knijff, and Manfred Kayser. Evaluating self-declared ancestry of us americans with autosomal, y-chromosomal and mitochondrial dna. *Human mutation*, 31(12):E1875–E1893, 2010.

009eye Fan Liu, Kate van Duijn, Johannes R Vingerling, Albert Hofman, André G Uitterlinden, A Cecile JW Janssens, and Manfred Kayser. Eye color and the prediction of complex phenotypes from genotypes. *Current Biology*, 19(5):R192–R193, 2009.

ncient Bastien Llamas, Lars Fehren-Schmitz, Guido Valverde, Julien Soubrier, Swapan Mallick, Nadin Rohland, Susanne Nordenfelt, Cristina Valdiosera, Stephen M Richards, Adam Rohrlach, et al. Ancient mitochondrial dna provides high-resolution time scale of the peopling of the americas. *Science advances*, 2(4):e1501385, 2016.

tabdpt Junwei Ma, Valentin Thomas, Rasa Hosseinzadeh, Hamidreza Kamkari, Alex Labach, Jesse C Cresswell, Keyvan Golestan, Guangwei Yu, Maksims Volkovs, and Anthony L Caterini. Tabdpt: Scaling tabular foundation models. *arXiv preprint arXiv:2410.18164*, 2024a.

ontext Junwei Ma, Valentin Thomas, Guangwei Yu, and Anthony Caterini. In-context data distillation with tabpfn. *arXiv preprint arXiv:2402.06971*, 2024b.

simons Swapan Mallick, Heng Li, Mark Lipson, Iain Mathieson, Melissa Gymrek, Fernando Racimo, Mengyao Zhao, Niru Chennagiri, Susanne Nordenfelt, Arti Tandon, et al. The simons genome diversity project: 300 genomes from 142 diverse populations. *Nature*, 538(7624):201–206, 2016.

ips23a D. McElfresh, S. Khandagale, J. Valverde, V. Prasad C, G. Ramakrishnan, M. Goldblum, and C. White. When do neural nets outperform boosted trees on tabular data? pages 76336–76369.

ssment Dennis McNevin, C Santos, A Gómez-Tato, J Álvarez-Dios, M Casares de Cal, R Daniel, C Phillips, and MV Lareu. An assessment of bayesian and multinomial logistic regression classification systems to analyse admixed individuals. *Forensic Science International: Genetics Supplement Series*, 4(1):e63–e64, 2013.

gender David I Min, Miyoung Lee, Yi-Min Ku, and Michael Flanigan. Gender-dependent racial difference in disposition of cyclosporine among healthy african american and white volunteers. *Clinical Pharmacology & Therapeutics*, 68(5):478–486, 2000.

enomic Luca Pagani, Daniel John Lawson, Evelyn Jagoda, Alexander Mörseburg, Anders Eriksson, Mario Mitt, Florian Clemente, Georgi Hudjashov, Michael DeGiorgio, Lauri Saag, et al. Genomic analyses inform on migration events during the peopling of eurasia. *Nature*, 538 (7624):238–242, 2016.

-learn F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. 12:2825–2830, 2011.

er2020 Peter Pfaffelhuber, Franziska Grundner-Culemann, Veronika Lipphardt, and Franz Baumdicker. How to choose sets of ancestry informative markers: A supervised feature selection approach. *Forensic Science International: Genetics*, 46:102259, 2020a.

choose Peter Pfaffelhuber, Franziska Grundner-Culemann, Veronika Lipphardt, and Franz Baumdicker. How to choose sets of ancestry informative markers: A supervised feature selection approach. *Forensic Science International: Genetics*, 46:102259, 2020b.

iaplex Chistopher Phillips, A Freire Aradas, AK Kriegel, M Fondevila, O Bulbul, C Santos, F Serrulla Rech, MD Perez Carceles, Á Carracedo, PM Schneider, et al. Eurasiaplex: a forensic snp assay for differentiating european and south asian ancestries. *Forensic Science International: Genetics*, 7(3):359–366, 2013.

NIPPER Chris Phillips, Antonio Salas, JJ Sanchez, Manuel Fondevila, Antonio Gómez-Tato, J Alvarez-Dios, Manuel Calaza, M Casares de Cal, David Ballard, María Victoria Lareu, et al. Inferring ancestral origin using a single multiplex assay of ancestry-informative marker snps. *Forensic Science International: Genetics*, 1(3-4):273–280, 2007a.

erring Chris Phillips, Antonio Salas, JJ Sanchez, Manuel Fondevila, Antonio Gómez-Tato, J Alvarez-Dios, Manuel Calaza, M Casares de Cal, David Ballard, María Victoria Lareu, et al. Inferring ancestral origin using a single multiplex assay of ancestry-informative marker snps. *Forensic Science International: Genetics*, 1(3-4):273–280, 2007b.

cestry Christopher Phillips, Lourdes Prieto, Manuel Fondevila, Antonio Salas, Antonio Gómez-Tato, José Álvarez-Dios, Antonio Alonso, Alejandro Blanco-Verea, María Brión, Marta Montesino, et al. Ancestry analysis in the 11-m madrid bomb attack investigation. *PloS one*, 4(8):e6583, 2009.

ps2014 Christopher Phillips, W Parson, B Lundsberg, C Santos, A Freire-Aradas, M Torres, M Eduardoff, C Børsting, P Johansen, M Fondevila, et al. Building a forensic ancestry panel from the ground up: The euroforgen global aim-snp set. *Forensic Science International: Genetics*, 11:13–25, 2014.

rd2000 Jonathan K Pritchard, Matthew Stephens, and Peter Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959, 2000.

ips18a L. Prokhorenkova, G. Gusev, A. Vorobev, A. Dorogush, and A. Gulin. Catboost: Unbiased boosting with categorical features. page 6639–6649.

98case Foster J Provost, Tom Fawcett, Ron Kohavi, et al. The case against accuracy estimation for comparing induction algorithms. In *ICML*, volume 98, pages 445–453, 1998.

rative Peter Resutik, Simon Aeschbacher, Michael Krützen, Adelgunde Kratzer, Cordula Haas, Christopher Phillips, and Natasha Arora. Comparative evaluation of the maplex, precision id ancestry panel, and visage basic tool for biogeographical ancestry inference. *Forensic Science International: Genetics*, 64:102850, 2023.

rg2003 Noah A Rosenberg, Lei M Li, Ryk Ward, and Jonathan K Pritchard. Informativeness of genetic markers for inference of ancestry. *The American Journal of Human Genetics*, 73(6): 1402–1422, 2003.

opment Jorge Ruiz-Ramírez, M de la Puente, Catarina Xavier, Adrián Ambroa-Conde, J Álvarez-Dios, A Freire-Aradas, Ana Mosquera-Miguel, Arwin Ralf, Christina Amory, Maria Alexandra Katsara, et al. Development and evaluations of the ancestry informative markers of the visage enhanced tool for appearance and ancestry. *Forensic Science International: Genetics*, 64:102853, 2023.

fiplex Carla Santos, Christopher Phillips, Manuel Fondevila, Runa Daniel, Roland AH Van Oorschot, Esteban G Burchard, Moses S Schanfield, Luis Souto, Jolame Uacyisrael, Marc Via, et al. Pacifiplex: an ancestry-informative snp panel centred on australia and the pacific region. *Forensic Science International: Genetics*, 20:71–80, 2016.

5nocit Harish Swaminathan, Catherine M Grgicak, Muriel Medard, and Desmond S Lun. Nocit: a computational method to infer the number of contributors to dna samples analyzed by str genotyping. *Forensic Science International: Genetics*, 16:172–180, 2015.

mental Vladimir Vovk. The fundamental nature of the log loss function. *Fields of logic and computation II: Essays dedicated To Yuri Gurevich on the Occasion of His 75th Birthday*, pages 307–318, 2015.

isplex Susan Walsh, Fan Liu, Kaye N Ballantyne, Mannis van Oven, Oscar Lao, and Manfred Kayser. Irisplex: a sensitive dna tool for accurate prediction of blue and brown eye colour in the absence of ancestry information. *Forensic Science International: Genetics*, 5(3):170–180, 2011.

isplex Susan Walsh, Fan Liu, Andreas Wollstein, Leda Kovatsi, Arwin Ralf, Agnieszka Kosiniak-Kamysz, Wojciech Branicki, and Manfred Kayser. The hirisplex system for simultaneous prediction of hair and eye colour from dna. *Forensic Science International: Genetics*, 7(1): 98–115, 2013.

enetic Sijia Wang, Cecil M Lewis Jr, Mattias Jakobsson, Sohini Ramachandran, Nicolas Ray, Gabriel Bedoya, Winston Rojas, Maria V Parra, Julio A Molina, Carla Gallo, et al. Genetic variation and population structure in native americans. *PLoS genetics*, 3(11):e185, 2007.

019lei Michael J Wathen, Yadu Gautam, Sudhir Ghandikota, Marepalli B Rao, and Tesfaye B Mersha. Lei: a novel allele frequency-based feature selection method for multi-ancestry admixed populations. *Scientific reports*, 9(1):11103, 2019.

rensic Yufeng Wen, Jing Liu, Yonglin Su, Xiacan Chen, Yiping Hou, Linchuan Liao, and Zheng Wang. Forensic biogeographical ancestry inference: recent insights and current trends. *Genes & genomics*, 45(10):1229–1238, 2023.

opment Catarina Xavier, Maria de la Puente, Ana Mosquera-Miguel, Ana Freire-Aradas, Vivian Kalamara, Athina Vidaki, Theresa E Gross, Andrew Revoir, Ewelina Pośpiech, Ewa Kartasińska, et al. Development and validation of the visage ampliseq basic tool to predict appearance and ancestry from dna. *Forensic Science International: Genetics*, 48:102336, 2020.

er2022 Catarina Xavier, Maria de la Puente, Maja Sidstedt, Klara Junker, Angelika Minawi, Martina Unterländer, Yann Chantrel, François-Xavier Laurent, Anna Delest, Carsten Hohoff, et al. Evaluation of the visage basic tool for appearance and ancestry inference using forenseq® chemistry on the miseq fgx® system. *Forensic Science International: Genetics*, 58:102675, 2022.

abflex Yuchen Zeng, Wonjun Kang, and Andreas C Mueller. Tabflex: Scaling tabular learning to millions with linear attention. In *NeurIPS 2024 Third Table Representation Learning Workshop*, 2024.