Exploring the complexity of African populations variability with Machine Learning

Tommaso Mori
Department of Biology
University of Florence
Firenze, Italy
tommaso.mori@unifi.it

Alessandro Riga
Department of Biology
University of Florence
Firenze, Italy
alessandro.riga@unifi.it

Jacopo Moggi-Cecchi
Department of Biology
University of Florence
Firenze, Italy
iacopo.moggicecchi@unifi.it

Chiara Canfailla
Department of Information Engineering
University of Florence
Firenze, Italy
chiara.canfailla@stud.unifi.it

Andrea Barucci
Institute of Applied Physics "Nello Carrara"
Italian National Research Council
Sesto Fiorentino, Italy
a.barucci@ifac.cnr.it

2022 IEEE Workshop on Complexity in Engineering (COMPENG) | 978-1-7281-7124-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/COMPENG50184.2022.9905451

***Abstract*—Human skeletal remains are an immense source of data to describe human biodiversity with an intrinsic complexity due to the multifactorial origin of human variability. Evolution and ontogeny produced complex patterns of variation through contingent events and adaptations. Multivariate approaches have been widely adopted in physical anthropology; however, at present, Artificial Intelligence algorithms have scarcely been applied to such datasets. Data analysis techniques based on Artificial Intelligence algorithms have shown to be suitable in many different fields, from engineering and medicine up to cultural heritage and Egyptology. In this work we aim to show how Machine Learning algorithms can be applied in the field of anthropology, using the W.W. Howells dataset of cranial measurements, limited to the analysis of African populations. Principal Component Analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE), Spectral Embedding and Uniform Manifold Approximation and Projection (UMAP) were used for dimensionality reduction, along with supervised and unsupervised methods to explore and quantify the differences due to ancestry and sex in the skulls of African populations. Algorithms such as Support Vector Machines and the unsupervised DBSCAN were applied to the data in order to quantify this similarity. This strategy allows a discrimination of sex and ancestry (about 85% of accuracy for both) in human remains, ultimately opening up new routes for anthropological research.**

***Index Terms*—component, formatting, style, styling, insert**

I. Introduction

Human skeletal remains provide a vast amount of data to study human variability. Extant human variability derives from multiple factors. Evolutionary history (in the generations) and ontogeny (in an individual's life course) shape the space of human variation, through contingent events and adaptations. The sources of variation are multiple and include sex and ancestry [1], [2] as well as cultural practices, lifestyle and socioeconomic conditions [3], [4]. The interplay between all these factors, produces complex patterns of variation, often difficult to disentangle.

However, thanks to the paradigm of Artificial Intelligence [5] (AI), newer tools are today available to explore this kind of data.

Machine Learning (ML) algorithms, in particular, are the core of many successful applications, covering nowadays every field of knowledge, from engineering [6], physics [7] and nanophotonics [8]–[10], up to medicine [11]–[14] and cultural heritage [15], [16].

Among the factors of variation, sex and ancestry are particularly important for forensic anthropologists, as they help in the identification of crime victims. However, the intrinsic complexity of human phenotype (and its continuous variation among sexes and populations) make it difficult to estimate those traits from metric data.

In this work we tried to disentangle such complexity using ML, employing the Howells dataset [17]–[19] of cranial measurements, in our application limited to the African populations. The idea is to identify a ML data analysis workflow, starting from data visualization up to supervised and unsupervised classification, supporting the work of anthropologists to identify sex and ancestry from human cranial remains. Recent works support our vision, highlighting the importance to introduce ML in Anthropology [20]–[24].

II. Workflow

In this work the following steps were performed:

- Dataset selection
- Machine Learning data analysis
- Anthropological interpretation of the results

A. Dataset selection - Cranial data

We selected the public available dataset [25] compiled by Dr. William Howells, comprising more than 3000 individuals

and up to 82 linear cranial measurements with notes about the provenance and sex of the individuals.

Aiming to challenge the discrimination power of ML algorithms, 5 African populations (Bushman[1], Dogon, Egyptian, Teita and Zulu) were selected.

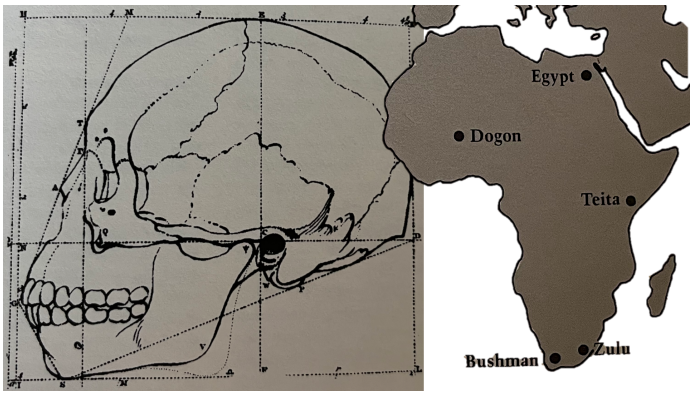

Fig. 1. Cranial example of measurements taken from Howells. On the right the geographic distribution of the 5 African populations.

B. Machine Learning

ML can be applied using a plethora of algorithms, some suitable to create comprehensive maps to visualize data distribution in 2D, other focused on classification of the data in a supervised or unsupervised way. However, due to the "No Free Lunch Theorem" [26], there is no way *a priori* to know which algorithm will perform better on our problem. So, the only solution is to try different algorithms in our dataset. In this work ML algorithms [27], [28] were used in order to explore the data to uncover sex and ancestry.

1) Dimensionality reduction techniques: Dimensionality reduction can be fulfilled with many algorithms, such as PCA, t-SNE [29]–[31], Self-Organizing map [32], and UMAP [33], [34]. In this work we started from the well-known PCA, its components potentially useful as input of the other algorithms, moving then to probabilistic techniques such as t-SNE [30], [31], UMAP (based on topological data analysis) [34] and Self Organizing map [32].

2) Supervised and unsupervised classification: Support Vector Machines [35], Random Forest (RF) [36], Neural Network (1 hidden layer allocating 100 neurons with ReLu as activation function, Adam solver and a regularization factor of 10^{-4}), k-Nearest Neighbor (kNN) [37] and AdaBoost [38] were used for supervised learning, while for unsupervised K-Means and DBSCAN [39], [40] were applied.

Cross validation was implemented to evaluate every performance of ML classification algorithms.

III. Results

A. Dimensionality reduction

PCA decomposition seems to suggest 5 clusters (Fig. 2). However, it is worth to note that in order to obtain an explained variance of about 80% 14 PC components must be taken into consideration.

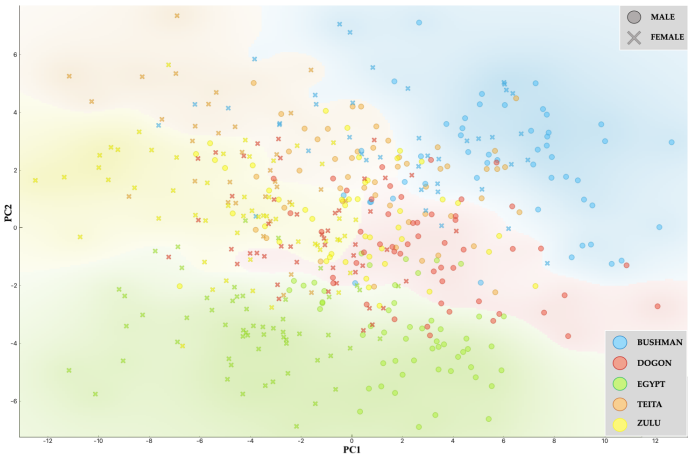

Fig. 2. PCA plots showing the first 2 components for all African populations.

Clearer results can be obtained using t-SNE (Fig. 3). Sex difference can introduces more variability, as observed by the difference in the 2D map between Female and Male (Fig. 4).

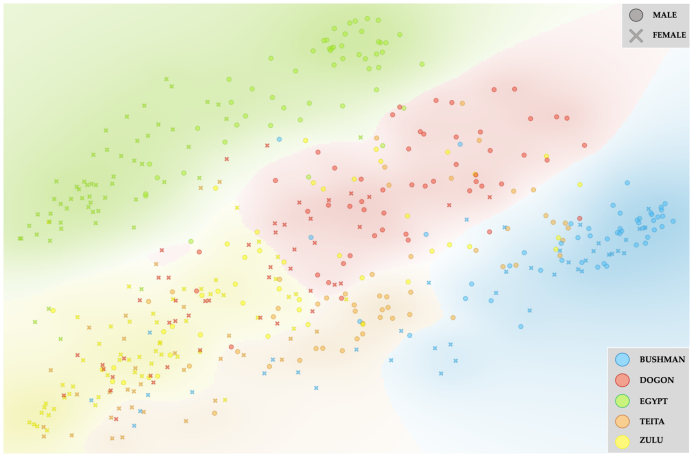

Fig. 3. Ancestry t-SNE results. Hyperparameters: Perp. 10 and Exag. 4.

Better results can then be obtained by restricting to just one sex, as shown in Fig. 5.

An example of dimensionality reduction, using UMAP applied to the Ancestry of African men, is reported in Fig. 7, while an example using Self Organizing Map applied to the Ancestry of African is illustrated in Fig. 8. All the algorithms suggest the same conclusions: Bushman and Egyptian are separated clusters, while the other three populations are more mixed.

B. Classification

Supervised classification results (average over all classes) are reported in Tab. I for Ancestry considering Male and Female or separately (Tab. II and Tab. III). Tab. IV reports results for Sex discrimination.

[1]We kept the Bushman wording, instead of San, as used in the original Howells database.

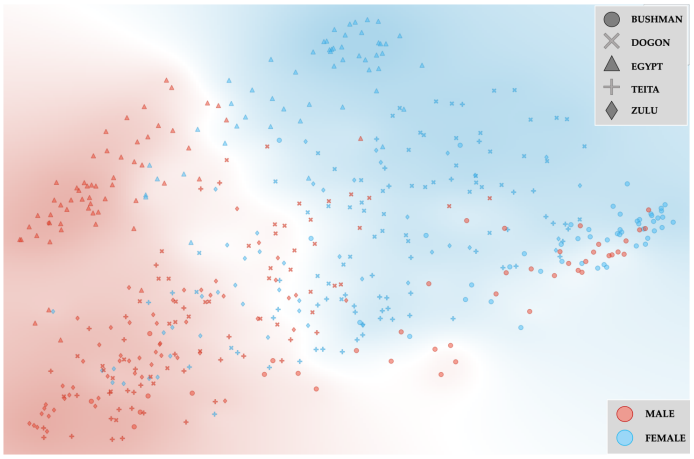

Fig. 4. Sex t-SNE results. Hyperparameters: Perp. 20 and Exag. 4.

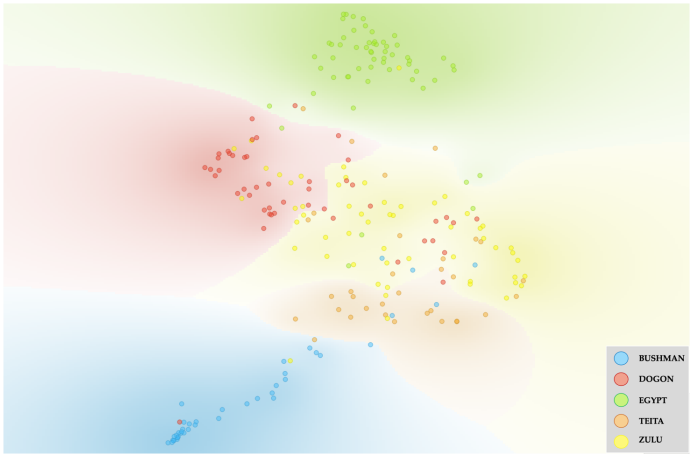

Fig. 5. Ancestry t-SNE results, restricting the analysis to Men only. Hyperparameters: Perp. 15 and Exag. 3.

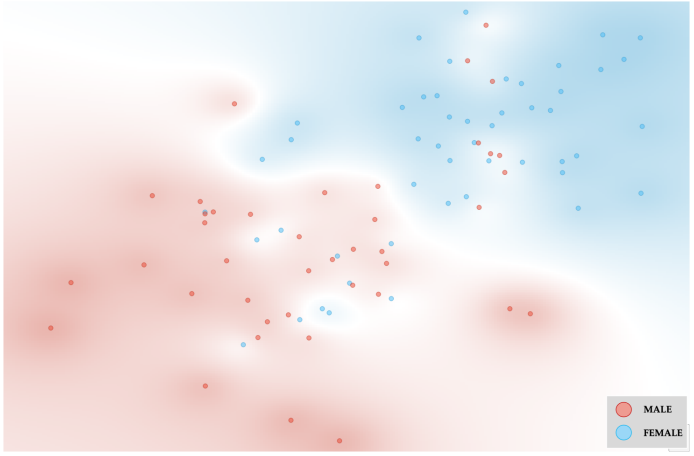

Fig. 6. Bushman Sex t-SNE results. Hyperparameters: Perp. 10 and Exag. 4.

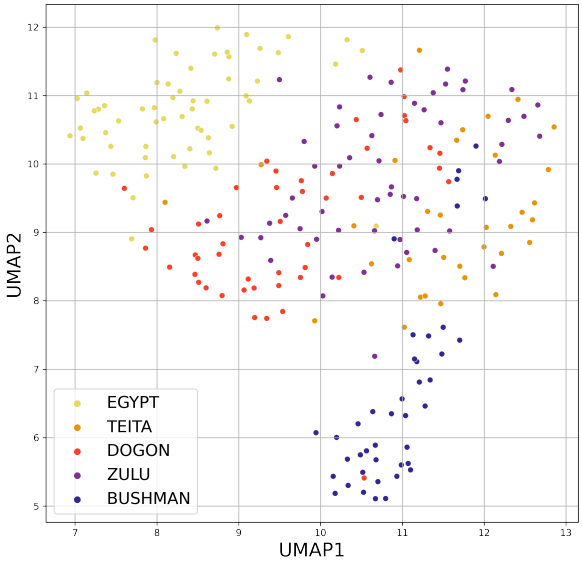

Fig. 7. UMAP Ancestry clustering of African men populations. Hyperparameters: number of neighbor = 30, minimum distance = 0.25.

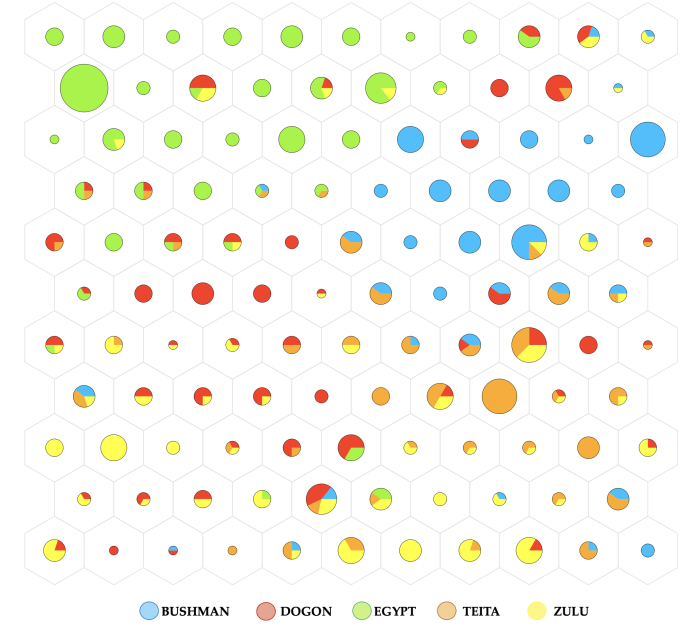

Fig. 8. Self Organizing map showing populations Ancestry clustering.

TABLE I
ANCESTRY SUPERVISED CLASSIFICATION RESULTS.

Algorithm	AUC	Accuracy	F1	Precision	Recall
SVM	0.98	0.87	0.87	0.87	0.87
RF	0.92	0.75	0.75	0.75	0.75
NN	0.985	0.89	0.89	0.89	0.89
kNN	0.93	0.76	0.76	0.76	0.76
AdaBoost	0.73	0.58	0.58	0.58	0.58

TABLE II
MEN ANCESTRY SUPERVISED CLASSIFICATION RESULTS.

Algorithm	AUC	Accuracy	F1	Precision	Recall
SVM	0.97	0.85	0.85	0.85	0.85
Random Forest	0.92	0.75	0.745	0.745	0.75
Neural Network	0.98	0.87	0.87	0.87	0.87
kNN	0.91	0.77	0.765	0.765	0.77
AdaBoost	0.77	0.64	0.64	0.64	0.64

TABLE III
FEMALE ANCESTRY SUPERVISED CLASSIFICATION RESULTS.

Algorithm	AUC	Accuracy	F1	Precision	Recall
SVM	0.97	0.87	0.87	0.87	0.87
Random Forest	0.94	0.75	0.745	0.745	0.75
Neural Network	0.98	0.88	0.88	0.89	0.88
kNN	0.93	0.75	0.74	0.75	0.75
AdaBoost	0.72	0.56	0.56	0.565	0.56

TABLE IV
SEX SUPERVISED CLASSIFICATION RESULTS FOR ALL POPULATIONS.

Algorithm	AUC	Accuracy	F1	Precision	Recall
SVM	0.93	0.86	0.86	0.86	0.86
Random Forest	0.91	0.83	0.83	0.83	0.83
Neural Network	0.93	0.85	0.85	0.85	0.85
kNN	0.88	0.79	0.79	0.80	0.79
AdaBoost	0.75	0.75	0.75	0.75	0.75

TABLE V
SEX SUPERVISED CLASSIFICATION RESULTS FOR BUSHMAN.

Algorithm	AUC	Accuracy	F1	Precision	Recall
SVM	0.82	0.78	0.78	0.78	0.78
Random Forest	0.83	0.73	0.73	0.73	0.73
Neural Network	0.85	0.77	0.77	0.77	0.77
kNN	0.765	0.70	0.70	0.70	0.70
AdaBoost	0.70	0.70	0.70	0.705	0.70

Results for unsupervised methods such as K-Means (Tab. VI) and DB-SCAN suggest 2 clusters relative to Sex, while Ancestry is complicated to uncover.

TABLE VI
K-MEANS RESULTS FOR ANCESTRY AND SEX.

Dataset	Sex	Objective	Clusters	Silhouette
All Populations	M+F	Ancestry	2	0.160
All Populations	Female	Ancestry	2	0.125
All Populations	Male	Ancestry	2	0.129
Bushman	M+F	Sex	2	0.166
Dogon	M+F	Sex	2	0.134
Egypt	M+F	Sex	2	0.174
Teita	M+F	Sex	2	0.154
Zulu	M+F	Sex	2	0.144

IV. DISCUSSION

Dimensionality reduction algorithms such as t-SNE and UMAP allow making some considerations on the relationships among the analysed populations. Egyptians are the only non sub-Saharan population; their separation from the others is likely linked to a reduced genic flow with sub-Saharan groups. Also the Bushman cluster separately, in agreement with their deep rooting in the phylogenetic tree of human populations [41].

It is worth noting as the importance of dimensionality reduction techniques must be researched not just in their power to visualize distinct clusters of data for a particular combination of hyperparameters, but rather in a wider view. E.g. Mixing or partial superimposition of clusters can give valuable information. Tweaking the hyperparameters such as metric, perplexity, and exaggeration for t-SNE, or the number of neighbors, minimum distance, number of components, and metric for UMAP, allows to obtain different visualization which must be carefully analyzed in order to extract meaningful knowledge [29], [33].

Classification results employing supervised learning algorithms are very promising, e.g. some of them (SVM and Neural Network) reach accuracy above 85% in ancestry and sex determination.

However, unsupervised classification unfortunately is not performing equally well. K-Means and DBscan don't allow discrimination between Ancestry always pinpointing to 2 clusters, while for Sex things seem better, but strong results are far away. These conclusions are reflected in the low values of the Silhouette score index [42], found for different combinations of Populations and Sex variables. For K-Means this can be ascribed to its ability to deal just with spherical clusters, which is not the case for our dataset. Surely, unsupervised methods deserve to be explored in details in future works, extending the kind of algorithms.

V. CONCLUSION

In this work we have shown how the problems of Ancestry and Sex determination in forensic anthropology can be tackled using machine learning algorithms, getting some understanding of the data structure, aiming to support the work of forensic anthropologists. Interesting results were found using dimensionality reduction algorithms and supervised approaches, while unsupervised methods deserve deeper investigation.

ACKNOWLEDGMENT

We acknowledge the support of Regione Toscana and the Sistema Museale di Ateneo for the funding of the project DIVINA (UNIFI FSC2022).

REFERENCES

[1] A. Del Bove, A. Profico, A. Riga, A. Bucchi, and C. Lorenzo, "A geometric morphometric approach to the study of sexual dimorphism in the modern human frontal bone," *American Journal of Physical Anthropology*, vol. 173, no. 4, pp. 643–654, 2020.

[2] E. A. DiGangi and J. T. Hefner, "Ancestry estimation," *Research methods in human skeletal biology*, pp. 117–149, 2013.

[3] T. Mori, A. Riga, G. Dionisio, F. Bigoni, and J. Moggi-Cecchi, "Cranial modification and trepanation in pre-hispanic collections from peru in the museum of anthropology and ethnology, florence, italy," *Medicina*, vol. 6, no. 1, p. e2022002, 2022.

[4] B. M. Holt and V. Formicola, "Hunters of the ice age: the biology of upper paleolithic people," *American Journal of Physical Anthropology: The Official Publication of the American Association of Physical Anthropologists*, vol. 137, no. S47, pp. 70–99, 2008.

[5] R. R. Dunn, M. C. Spiros, K. R. Kamnikar, A. M. Plemons, and J. T. Hefner, "Ancestry estimation in forensic anthropology: A review," *Wiley Interdisciplinary Reviews: Forensic Science*, vol. 2, no. 4, p. e1369, 2020.

[6] I. H. Sarker, "Machine learning: Algorithms, real-world applications and research directions," *SN Computer Science*, vol. 2, no. 3, pp. 1–21, 2021.

[7] G. Carleo, I. Cirac, K. Cranmer, L. Daudet, M. Schuld, N. Tishby, L. Vogt-Maranto, and L. Zdeborová, "Machine learning and the physical sciences," *Reviews of Modern Physics*, vol. 91, no. 4, p. 045002, 2019.

[8] C. Borri, S. Centi, S. Chioccioli, P. Bogani, F. Micheletti, M. Gai, P. Grandi, S. Laschi, F. Tona, A. Barucci *et al.*, "based genetic assays with bioconjugated gold nanorods and an automated readout pipeline," *Scientific Reports*, vol. 12, no. 1, pp. 1–12, 2022.

[9] S. Guo, J. Popp, and T. Bocklitz, "Chemometric analysis in raman spectroscopy from experimental design to machine learning–based modeling," *Nature protocols*, vol. 16, no. 12, pp. 5426–5459, 2021.

[10] A. Barucci, C. D'Andrea, E. Farnesi, M. Banchelli, C. Amicucci, M. de Angelis, B. Hwang, and P. Matteini, "Label-free sers detection of proteins based on machine learning classification of chemo-structural determinants," *Analyst*, vol. 146, no. 2, pp. 674–682, 2021.

[11] O. Koteluk, A. Wartecki, S. Mazurek, I. Kołodziejczak, and A. Mackiewicz, "How do machines learn? artificial intelligence as a new era in medicine," *Journal of Personalized Medicine*, vol. 11, no. 1, p. 32, 2021.

[12] B. Bhinder, C. Gilvary, N. S. Madhukar, and O. Elemento, "Artificial intelligence in cancer research and precision medicine," *Cancer discovery*, vol. 11, no. 4, pp. 900–915, 2021.

[13] M. Avanzo, M. Porzio, L. Lorenzon, L. Milan, R. Sghedoni, G. Russo, R. Massafra, A. Fanizzi, A. Barucci, V. Ardu *et al.*, "Artificial intelligence applications in medical imaging: A review of the medical physics research in italy," *Physica Medica*, vol. 83, pp. 221–241, 2021.

[14] E. Bertelli, L. Mercatelli, C. Marzi, E. Pachetti, M. Baccini, A. Barucci, S. Colantonio, L. Gherardini, L. Lattavo, M. A. Pascali *et al.*, "Machine and deep learning prediction of prostate cancer aggressiveness using multiparametric mri," *Frontiers in oncology*, vol. 11, 2021.

[15] A. Barucci, C. Cucci, M. Franci, M. Loschiavo, and F. Argenti, "A deep learning approach to ancient egyptian hieroglyphs classification," *IEEE Access*, vol. 9, pp. 123 438–123 447, 2021.

[16] C. Cucci, A. Barucci, L. Stefani, M. Picollo, R. Jiménez-Garnica, and L. Fuster-Lopez, "Reflectance hyperspectral data processing on a set of picasso paintings: which algorithm provides what? a comparative analysis of multivariate, statistical and artificial intelligence methods," in *Optics for Arts, Architecture, and Archaeology VIII*, vol. 11784. International Society for Optics and Photonics, 2021, p. 1178404.

[17] W. W. Howells, "Cranial variation in man: a study by multivariate analysis of patterns of difference among recent human populations," *Peabody Museum of Archaeology and Ethnology, Harvard Univ.*, 1973.

[18] ——, "Skull shapes and the map: craniometric analyses in the dispersion of modern homo," *Papers of the Peabody museum of Archaeology and Ethnology*, vol. 79, 1989.

[19] ——, "Who's who in skulls: ethnic identification of crania from measurements," *Papers of the Peabody Museum of Archaeology and Ethnology*, vol. 82, 1995.

[20] E. Alladio, B. Poggiali, G. Cosenza, and E. Pilli, "Multivariate statistical approach and machine learning for the evaluation of biogeographical ancestry inference in the forensic field," *Scientific Reports*, vol. 12, no. 1, pp. 1–17, 2022.

[21] K. Sun, Y. Yao, L. Yun, C. Zhang, J. Xie, X. Qian, Q. Tang, and L. Sun, "Application of machine learning for ancestry inference using multi-indel markers," *Forensic Science International: Genetics*, vol. 59, p. 102702, 2022.

[22] A. Budiarto and B. Pardamean, "Explainable supervised method for genetics ancestry estimation," in *2021 1st International Conference on Computer Science and Artificial Intelligence (ICCSAI)*, vol. 1. IEEE, 2021, pp. 422–426.

[23] Y. Dong, A. Gao, I. Hou, K. Ma, R. Huang, Y. Bai, and X. Liu, "A deep learning model for ancestry estimation with craniometric measurements," in *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2021, pp. 3350–3357.

[24] T. Mori and K. Harvati, "Basicranial ontogeny comparison in pan troglodytes and homo sapiens and its use for developmental stage definition of knm-er 42700," *American Journal of Physical Anthropology*, vol. 170, no. 4, pp. 579–594, 2019.

[25] D. B. M. Auerbach. William w. howells craniometric data set. [Online]. Available: https://web.utk.edu/ auerbach/HOWL.htm

[26] D. H. Wolpert and W. G. Macready, "No free lunch theorems for optimization," *IEEE transactions on evolutionary computation*, vol. 1, no. 1, pp. 67–82, 1997.

[27] J. Demšar, T. Curk, A. Erjavec, Črt Gorup, T. Hočevar, M. Milutinovič, M. Možina, M. Polajnar, M. Toplak, A. Starič, M. Štajdohar, L. Umek, L. Žagar, J. Žbontar, M. Žitnik, and B. Zupan, "Orange: Data mining toolbox in python," *Journal of Machine Learning Research*, vol. 14, pp. 2349–2353, 2013. [Online]. Available: http://jmlr.org/papers/v14/demsar13a.html

[28] 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," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[29] D. Kobak and P. Berens, "The art of using t-sne for single-cell transcriptomics," *Nature communications*, vol. 10, no. 1, pp. 1–14, 2019.

[30] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.

[31] M. Wattenberg, F. Viégas, and I. Johnson, "How to use t-sne effectively," *Distill*, 2016. [Online]. Available: http://distill.pub/2016/misread-tsne

[32] T. Kohonen, "Essentials of the self-organizing map," *Neural networks*, vol. 37, pp. 52–65, 2013.

[33] E. Becht, L. McInnes, J. Healy, C.-A. Dutertre, I. W. Kwok, L. G. Ng, F. Ginhoux, and E. W. Newell, "Dimensionality reduction for visualizing single-cell data using umap," *Nature biotechnology*, vol. 37, no. 1, pp. 38–44, 2019.

[34] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2018.

[35] J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, and A. Lopez, "A comprehensive survey on support vector machine classification: Applications, challenges and trends," *Neurocomputing*, vol. 408, pp. 189–215, 2020.

[36] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[37] A. Mucherino, P. J. Papajorgji, and P. M. Pardalos, "K-nearest neighbor classification," in *Data mining in agriculture*. Springer, 2009, pp. 83–106.

[38] R. E. Schapire, "Explaining adaboost," in *Empirical inference*. Springer, 2013, pp. 37–52.

[39] M. Ester, H.-P. Kriegel, J. Sander, X. Xu *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise." in *kdd*, vol. 96, no. 34, 1996, pp. 226–231.

[40] E. Schubert, J. Sander, M. Ester, H. P. Kriegel, and X. Xu, "Dbscan revisited, revisited: why and how you should (still) use dbscan," *ACM Transactions on Database Systems (TODS)*, vol. 42, no. 3, pp. 1–21, 2017.

[41] C. M. Schlebusch, P. Sjödin, G. Breton, T. Günther, T. Naidoo, N. Hollfelder, A. E. Sjöstrand, J. Xu, L. M. Gattepaille, M. Vicente, D. G. Scofield, H. Malmström, M. de Jongh, M. Lombard, H. Soodyall, and M. Jakobsson, "Khoe-San Genomes Reveal Unique Variation and Confirm the Deepest Population Divergence in Homo sapiens," *Molecular Biology and Evolution*, vol. 37, no. 10, pp. 2944–2954, 07 2020. [Online]. Available: https://doi.org/10.1093/molbev/msaa140

[42] K. R. Shahapure and C. Nicholas, "Cluster quality analysis using silhouette score," in *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, 2020, pp. 747–748.