Bioarchaeological sex prediction from central Italy using generalized low rank imputation for cross-validated metric craniodental supervised ensemble machine learning with missing data

Author: Evan Muzzall (evan.muzzall@berkeley.edu)

Institutional affiliation: D-Lab, University of California Berkeley

keywords: Sex estimation, SuperLearner ensemble machine learning, cross validation, generalized low rank model, central Italy

Abstract

I use a novel supervised ensemble machine learning approach to verify sex estimation of archaeological skeletons from central Italian bioarchaeological contexts with large amounts of missing data present. Eighteen cranial interlandmark distances and five maxillary metric distances were recorded from *n* = 240 estimated males and *n* = 180 estimated females from four locations at Alfedena (600-400 BCE) and two locations at Campovalano (750-200 BCE and 9-11th Century CE). A generalized low rank model (GLRM) was used to impute missing data and 20-fold external stratified cross-validation was used to fit an ensemble of eight machine learning algorithms to six different subsets of the data: 1) the face, 2) vault, 3) cranial base, 4) combined face/vault/base, 5) dentition, and 6) combined cranianiodental. Area under the receiver operator characteristic curve (AUC) was used to evaluate the predictive performance of six constituent algorithms, the discrete algorithmic winner(s), and the SuperLearner weighted ensemble's classification of males and females from these six bony regions. This approach is useful for predicting male/female sex from central Italy. AUC for the combined craniodental data was the highest (0.9722), followed by the combined cranial data (0.9644), the face (0.9426), vault (0.9116), base (0.9060), and dentition (0.7421). Cross-validated ensemble machine learning of cranial and dental data shows strong potential for estimating sex in the bioarchaeological record and can contribute additional perspectives to help refine our understanding of human sex estimation. Additionally, GLRMs have the potential to handle missing data in ways previously unexplored in the discipline. The main limitation is that the biological sexes of the individuals estimated in this study are not certain, but were estimated macroscopically using common bioarchaeological methods. However, these methods show great promise for estimation of sex in bioarchaeological and forensic contexts and should be investigated on known-sex reference samples for confirmation.

Introduction

Accurate sex estimation of archaeological skeletal remains is a fundamental step for reconstructing biological and demographic profiles of past humans. After unknown remains are identified, documented, and recovered, the sex and age of deceased individuals and groups are commonly estimated using traditional methods of measuring the pelvis, skull, and teeth (Buikstra & Ubelaker, 1994; Garvin & Ruff, 2012; Krishan et al., 2016). However, because female and male biological maturation rates differ (Slemenda et al., 1994; Wang, 2002), sex misidentification can lead to data recording bias and depreciated interpretability. After sex has been estimated and with the assistance of other biological and archaeological contextual information, the identities and lifeways of the deceased can be reconstructed in bioarchaeological contexts. However, traditional macroscopic sexing methods possess varying degrees of accuracy (Weiss, 1972; Sutter, 2003; Konigsberg, Algee-Hewitt, & Steadman, 2009; Jackes, 2011; Sierp & Henneberg, 2015; Irurita Olivares & Alemán Aguilera, 2016). For example, the dentition, pelvis, and crania can provide different results for estimating sex of the deceased. Tooth crown calcification and eruption and bone epiphyseal fusion are useful until early adulthood when 3rd molars erupt and bony ossification centers fuse in their final, united shapes. However, pelvic and cranial suture methods are used to estimate age in individuals through later stages of adulthood, albeit with wider margins of error.

Craniometric dimensions are frequently used as proxies for genetic relatedness due to their potentially heritable nature and correlations with neutral and adaptive genetic variation and selection (Sjøvold, 1984; Devor, 1987; Roseman, 2004; Roseman & Weaver, 2004; Carson, 2006; Witherspoon et al., 2007; Martínez-Abadías et al., 2009; Strauss & Hubbe, 2010; Herrera, Hanihara, & Godde, 2014). In the absence of genetic information, these methods are used to approximate the genetic and evolutionary relationships of past humans (Buikstra, Frankenberg, & Konigsberg, 1990), thus making accurate sex classification an integral first step in the reconstruction of other biological and demographic parameters. Hence, further examinations of sex correlations with other lines of evidence such as burial location, material culture, occupation, health, trauma prevalence, and biological relatedness will be skewed if sex is first misclassified.

Machine learning has yet to gain a foothold in bioarchaeology despite our discipline's deep ties to statistics and computational research for investigation of large quantitative datasets. Cunningham's (1997) pioneering machine learning social anthropological work for rule-based kinship structure detection set a high bar for anthropologists of all subdisciplines to aspire. However, her work remains largely unrecognized even though it exemplifies the types of problem-and-dataset-driven questions faced by bioarchaeologists. This discrepancy persists despite the success of bioarchaeological machine learning applications for estimating sex, age, ancestry, body mass, and stature in forensic anthropology, radiography, and anatomy (Bell & Jantz, 2001; Hefner & Ousley, 2014; Czibula, Ionescu, Miholca, & Mircea, 2016; Ionescu, Teletin, & Voiculescu, 2016; Ionescu, Czibula, & Teletin, 2018; Miholca, Czibula, Mircea, Czibula, 2016; Pink, 2016; Porto et al., 2019; Ortiz, Costa, Silva, Biazevic, & Michel-Crosato, 2020). Even less bioarchaeological research has focused on missing data imputation (Kenyhercz & Passalacqua, 2016).

Therefore, more examples are needed to better test our methodological understandings of skeletal and dental sex estimation. This research is an extension of Muzzall, Kennedy, and Culich (2017) which improved sex prediction accuracy of Howells Worldwide Craniometric Dataset and provided another example of the strong potential for machine learning to assist in sex estimation in bioarchaeological contexts. Here, I use a generalized low rank model to impute large amounts of missing data for a cross-validated supervised ensemble machine learning approach. This framework consists of eight algorithms total and is fit to cranial interlandmark and dental metric distances to predict binary sex from six pelvic and cranially estimated samples at Alfedena (600-400 BCE) and Campovalano (750-200 BCE and 9-11th Centuries CE) in central Italy.

Italy is home to one of the most colossal bioarchaeological contexts on Earth and represents humans' deep history throughout the region. Its central Mediterranean location and long and complex temporal histories and geological and environmental diversity have been influential in shaping the genetic, morphological, and cultural histories of the region (Scozzari et al., 2001; Coppa, Cucina, Lucci, Mancinelli, & Vargiu, 2007; Muttoni, Scardia, Kent, Swisher, & Manzi, 2009; Fu, Rudan, Pääbo, & Krause, 2012; Ghirotto et al., 2013). Humans here developed some of the richest and most divergent forms of worship, architecture, iconography and writing, and empires that persisted for long periods of time and across the globe via trade, warfare, and colonization. Central Italy was a particular crossroads between Africa and Europe and the Near East and Iberia and was home to many chiefdoms and nation-states that contained both shared and varied forms of settlement patterns, social and burial organization, material cultures, mortuary behaviors, and skeletal-dental morphologies (Muzzall & Coppa, 2019). As a result, Italy's bioarchaeological record provides a space to experiment with new methodologies for sex estimation.

Materials and Methods

Data

The dataset consists of metric cranial and dental data from 240 males and 180 females from central Italy: four locations at the Iron Age necropolis at Alfedena (600-400 BCE), the Iron Age graveyard at Campovalano (750-200 BCE), and the Medieval cemetery at Campovalano (9-11th Centuries CE) (Table 1). Cranial metric data were collected from a total of twelve standard anatomical landmarks: four from the face, four from the cranial vault, and four from the cranial base (Table 2). This produced a total of eighteen cranial interlandmark distances, six from each of the four landmarks from the three cranial regions.

Dental metric data were comprised of maximum mesiodistal dimensions of the right (or left-substituted when right antimere was missing) maxillary canine (XC) and buccolingual breadths of the right mesial (P3) and distal (P4) premolars and first (M1) and second (M2) molars (Hillson et al., 2006). Thus, six different subsets of the data were used: 1) six metrics from the face, 2) six from the vault, 3) six from the base, 4) eighteen from the cranium (the combined face, vault, and base metrics), 5) five from the dentition, and 6) twenty-three from the total combined cranial and dental data. Tukey boxplots are used to illustrate sex differences in these metrics. Sex was originally estimated macroscopically for all samples by Coppa and Macchiarelli (1982) and Bondioli, Corruccini, & Macchiarelli (1986) using methods for the pelvis and crania, except for Campovalano St. Peter which was estimated by EM using discrete cranial traits (Buikstra and Ubelaker, 1994).

Table 1. Location, time period, and sex distributions for males and females from Central Italy used in this study.

Location	Time period	Female	Male
Alfedena Arboreto	600-400 BCE	10	9
Alfedena Campo Consolino	600-400 BCE	19	61
Alfedena Scavi Mariani	600-400 BCE	28	37
Alfedena Sergi Museum	600-400 BCE	13	19
Campovalano Iron Age	750-200 BCE	77	89
Campovalano St. Peter	9-11th C. CE	33	25
Total		180	240

Table 2. Cranial anatomical landmarks used in this study. The four landmarks from each of the three regions produced eighteen total interlandmark distances – six for each region.

Face	Definition
Nasion (n)	The intersection of the naso-frontal suture in the midsagittal plane.
Prosthion (pr)	The location of the anteriorly located portion of the anterior surface of the alveolar process at the most anterior point of the alveolar process
Right frontomalare orbitale (fmorR)	The location where the zygomaticofrontal suture intersects the orbital margin
Left zygomaxillare (zymL)	The most inferior and anterior location on the zygomaticomaxillary suture
Vault	
Bregma (b)	The landmark where the sagittal and coronal sutures meet in the midsagittal plane. In cases where the sagittal suture deflects laterally, an estimation must be made of the location in the midsagittal plane.
Lambda (l)	The landmark where the left and right lambdoidal sutures intersect the sagittal suture. The landmark must be estimated when the suture intersection is obliterated, or where strongly serrated sutures are present
Right Asterion (astR)	The juncture of the lambdoid, parietomastoid, and occipitomastoid sutures
Left Frontotemporale (ftL)	The most medial and anterior point on the superior temporal line on the frontal bone

Base	
Nasion (n)	The intersection of the naso-frontal suture in the midsagittal plane.
Basion (ba)	The inner border where the anterior portion of the foramen magnum is intersected by the midsagittal plane
Hormion (h)	The juncture of the sphenoid and vomer bones in the midsagittal plane
Left Porion (poL)	The most superior point on the external margin of the external auditory meatus

Missing data

Missing data were prevalent from all areas of measurement and percentages of missing values for the face, vault, base, and dentition are shown in Table 3. A generalized low rank model (GLRM) was used to impute the missing values and function as an extension of principal component analysis (PCA) for low rank matrix tabular dataset approximation, by

> "approximating a data set as a product of two low dimensional factors by minimizing an objective function. The objective will consist of a loss function on the approximation error together with regularization of the low dimensional factors. With these extensions of PCA, the resulting low rank representation of the data set still produces a low dimensional embedding of the data set, as in PCA" (Udell, Horn, Zadeh, & Boyd, 2016: 3)

A generalized low rank model is essentially an unsupervised approach for data completion that uses clustering of known data in reduced dimensional space. The advantage of this data-adaptive approach to reconstruct missingness in the skeletal and dental remains instead of column mean, median, or k-nearest neighbor imputation is that it effectively uses clustering of features to impute the missing data, which makes sense given that the missingness of the data arises directly from missingness in the remains. Missingness indicators were also added as columns to the dataset to indicate exactly where missing and imputed data were located. These columns also functioned as predictor variables in the machine learning models to see if the location of missing data was related to sex prediction ability.

Table 3. Percentage of missing data for each variable. Missing data were imputed via generalized low rank model.

Bony region	Measurement	% Missing Female	% Missing Male
Face	n_pr	0.67	0.63
	n_fmorR	0.58	0.54
	n_zymL	0.65	0.57
	pr_fmorR	0.68	0.63
	pr_zymL	0.69	0.63

	fmorR_zymL	0.71	0.63
Vault	b_l	0.47	0.38
	b_astR	0.46	0.38
	b_ftL	0.51	0.42
	l_astR	0.44	0.37
	l_ftL	0.54	0.44
	astR_ftL	0.54	0.46
Base	n_ba	0.66	0.61
	n_h	0.68	0.63
	n_poL	0.61	0.53
	ba_h	0.69	0.65
	ba_poL	0.62	0.57
	h_poL	0.66	0.61
Dentition	XC	0.69	0.59
	P3	0.63	0.53
	P4	0.66	0.50
	M1	0.46	0.49
	M2	0.53	0.53

Ensemble Machine learning

Machine learning is defined as "a vast set tools for understanding data" (James, Witten, Hastie, and Tibshirani, 2013:1). It originated as a combination of computer science and statistics, but its greatest strength is its breadth of research application (Breiman, 2001a; Welling, 2015). Early examples stem from the social and cognitive sciences that attempted to predict and imitate human behavior (Turing, 1950; Rosenblatt, 1958; Samuel, 1959). In this research I use a supervised classification machine learning approach because the goal is to predict a categorical outcome (binary male/female sex) using craniodental features as predictor variables.

Ensembles are useful supervised machine learning methods because they optimize predictor accuracy through combinations of a suite of less accurate models (Dietterich, 2000). The SuperLearner approach (van der Laan, Polley, & Hubbard, 2007; Polley & van der Laan, 2010) is an algorithm that uses cross-validation (Efron & Gong, 1982) to estimate the performance of several machine learning models, and/or the same algorithm(s) with different settings. It produces an optimal weighted average of those models (an "ensemble model"), using cross-validated performance. This approach is asymptotically as accurate as the best single prediction algorithm that is tested. I fit the same machine learning ensemble of the eight algorithms (six constituent algorithms, the weighted SuperLearner ensemble, and the single best “discrete” algorithm) to predict binary sex classification (male/female) for each of the six subsets of the data described above as the predictors: face, vault, base, combined cranial, dental, and combined craniodental.

Table 4. List of eight machine learning algorithms used in this research.

Algorithm	Description	Reference
Logistic regression	Logistic regression models the relationships between the outcome variable (male/female sex) and the predictor variables. It computes the probability that the Y variable (sex) belongs to one of the two binary classes.	Dobson, 1990
Lasso	Lasso (least absolute shrinkage and selection operator) is a form of penalized regression (L1) that produces a sparse solution to remove predictor variables from the model that are not related to the outcome.	Friedman et al,. 2010
Decision tree	A decision tree is a relatively simple tree-based method that gauges the probability of classifying the outcome based on the predictor variables before splitting a given decision node a certain number of times until there is no longer enough observations to split.	Breiman et al., 1984
Ranger (random forest)	Ranger is a decorrelated random forest ensemble classifier method that uses the average of multiple bootstrapped decision tree models for classification. Unlike single decision tree models that use all predictors at each split, random forests use only a random subsample of the total predictors for each split in each tree.	Breiman, 2001b; Wright and Ziegler, 2017
Xgboost	A gradient boosted machine is also a tree-based method that fits a tree to the residuals of the previous tree in succession. It downweights easily predicted cases but upweights those that it cannot predict. This continues over many iterations so that weak trees are "boosted" into strong ones.	Freund and Schapire, 1999; Chen et al,. 2019
SuperLearner	The SuperLearner algorithm is an optimal weighted ensemble average that improves predictor construction and is flexible in that it can perform well on different data distributions and protects against overfitting through external cross-validation. Individual algorithm weights can be investigated to see which ones contribute most to the ensemble.	van der Laan et al., 2007; Kennedy, 2017
DiscreteSL	The discrete SuperLearner – the single best performing algorithm. This might also correspond to the combination of best performing algorithms at different cross-validation folds, in which case the DiscreteSL AUC will not be identical to that of a single algorithm.	Polley and van der Laan, 2010
Mean of Y	The mean of Y is the benchmark algorithm based only on the mean. This is a very simple prediction so the more complex algorithms should perform better than	Polley and van der Laan, 2010

	the sample Y mean. It should not be the best single-performing algorithm and should have a low weight in the weighted-average ensemble. If it is the best algorithm something is likely wrong.

Evaluating model performance

Stratified cross-validated area under the receiver operator characteristic curve (AUC) was used to evaluate the performance of the individual algorithms and the weighted SuperLearner ensemble (Lantz, 2015; Kennedy, 2017). The receiver operator characteristic curve itself represents the probability that a binary outcome (female or male, in this case) is correctly classified (Hanley and McNeil, 1982) while the AUC provides the degree of separability for the sexes that the model achieves. The receiver operator characteristic curve models the sensitivity (true positive rate) versus specificity (true negative rate) at various thresholds along the receiver operator characteristic curve. Maximization of AUC is ideal, which ranges from 0.5 (equivalent to random guessing) to 1.0 (perfect prediction). AUC is more useful for prediction of imbalanced classes and to prevent overfitting of a single class compared to simple classification accuracy.

Instead of fitting the models separately and looking at the performance (lowest risk), algorithms should be fit simultaneously. Risk is the average loss function used here and measures how far off the prediction was for a given observation and is calculated by non-negative least squares error; the lower the risk the fewer errors were made by the model. SuperLearner also identifies which single algorithm (or combination of algorithms) is best (the "discrete winner"), in addition to calculating the weighted average of the ensemble itself. Coefficient weights can be viewed to see each algorithm's contribution to this weighted ensemble average.

Stratified *k*-fold cross-validation is a process that divides the data into equally sized portions and trains a model on *k*-1 portions of the data so that the model can learn the relationship between male/female sex and the various predictors. The one holdout portion is used for testing purposes (but not for fitting the SuperLearner) and this process is repeated *k* times. I chose 20 folds, so each algorithm was trained on 19 portions of the data (95%) and tested on the one holdout (5%), twenty times. This also produces standard errors for the performance of each algorithm that can be compared to the SuperLearner average. Analysis was conducted in R version 3.6.2 and the ck37r, SuperLearner, and ggplot2 packages (Wickham, 2016; Polley, LeDell E, Kennedy, & van der Laan, 2019; Kennedy, 2020).

Results

Results indicate that ensemble machine learning has strong potential for sex prediction and yielded AUC values greater than 0.90 for the cranial metric data and ~0.74 for the dental metric data, despite large amounts of missing data. Males are larger than females in all dimensions as shown by the Tukey boxplots in Figures 1 and 2 although distributions for the sexes overlap considerably.

AUC performance for each algorithm along with their standard errors and confidence intervals are shown in Table 5. The combined craniodental data had the highest AUC with 0.9722, followed by the combined cranial (0.9644), face (0.9426), vault (0.9116), base (0.9060), and dentition (0.7421). Expectedly, the mean of Y is the worst performing algorithm in all cases (AUC = 0.500 for each). The SuperLearner algorithm has the highest AUC for all six bony regions while ranger is a close second for the face, vault, base, cranial, and combined craniodental data. Logistic regression is a close second for the dental data.

Also, the single best algorithm (or combination of algorithms) – the DiscreteSL – for the combined craniodental data was the lasso algorithm, with it performing the best for all 20 external cross-validation folds. Ranger was the best-performing algorithm all 20 times for the face, base, and combined cranial data. However, for the vault, ranger was the best performing algorithm 19 times and the decision tree algorithm once. For the dental data, logistic regression was the best performing algorithm 14 times, lasso 4 times, and ranger twice – this algorithmic confusion could be related to the considerably lower AUC for the dentition compared to any of the cranial data.

The SuperLearner weight distributions show which of the individual algorithms contributed most to the ensemble (Table 6). For the combined craniodental data, lasso contributed a coefficient of 0.4522, indicating that it contributed this percentage to the SuperLearner ensemble. This was followed by lesser contributes from the ranger algorithm (0.1734), xgboost (0.1700), logistic regression (0.1319), and decision tree (0.0726). For cranial data, ranger contributed a coefficient of 0.4610, followed by lesser contributions from logistic regression (0.1940), lasso (0.1411), decision tree (0.1267), and xgboost (0.0772). Contributions to the face stem mostly from ranger (0.4634) and logistic logistic regression (0.4193), for the vault from ranger (0.5004) and decision tree (0.3234), and for the base from ranger (0.8878). For the dentition, contributions stem mostly from logistic regression (0.5591) and ranger (0.3582).

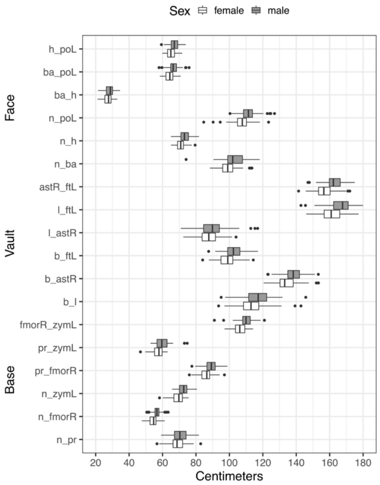

Figure 1. Distributions of raw cranial data for males and females. Anatomical landmark abbreviations are found in Table 2.

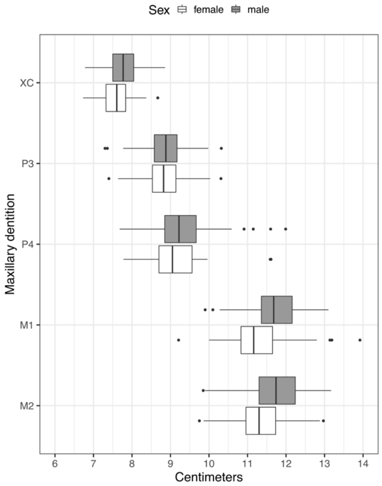

Figure 2. Distributions of raw dental data for males and females. Anatomical landmark abbreviations are found in Table 2.

Table 5. Cross-validated external AUC statistics for the six different measurement regions. 0.5 is the equivalent of random guessing; 1 means perfect prediction.

Bony region	Algorithm	AUC	Standard error	Confidence interval (lower)	Confidence interval (upper)
Face	Mean of Y	0.5000	0.0493	0.4034	0.5966
	Decision tree	0.8069	0.0259	0.7562	0.8577
	Xgboost	0.8998	0.0152	0.8701	0.9295
	Lasso	0.9042	0.0161	0.8727	0.9357
	Logistic regression	0.9088	0.0157	0.8781	0.9395
	Ranger	0.9306	0.0122	0.9066	0.9545
	DiscreteSL	0.9306	0.0122	0.9066	0.9545
	SuperLearner	0.9426	0.0111	0.9208	0.9644
Vault	Mean of Y	0.5000	0.0493	0.4034	0.5966
	Logistic regression	0.8458	0.0200	0.8067	0.8850
	Lasso	0.8486	0.0198	0.8099	0.8873
	Xgboost	0.8690	0.0188	0.8322	0.9058
	Decision tree	0.8998	0.0218	0.8570	0.9425
	DiscreteSL	0.9030	0.0164	0.8709	0.9351
	Ranger	0.9065	0.0158	0.8756	0.9374
	SuperLearner	0.9116	0.0147	0.8827	0.9404

Base	Mean of Y	0.5000	0.0493	0.4034	0.5966
	Logistic regression	0.7667	0.0238	0.7201	0.8132
	Lasso	0.7685	0.0238	0.7219	0.8152
	Decision tree	0.7986	0.0248	0.7500	0.8472
	Xgboost	0.8646	0.0177	0.8298	0.8993
	Ranger	0.9051	0.0146	0.8764	0.9338
	DiscreteSL	0.9051	0.0146	0.8764	0.9338
	SuperLearner	0.9060	0.0146	0.8774	0.9347
Cranial	Mean of Y	0.5000	0.0493	0.4034	0.5966
	Decision tree	0.9125	0.0189	0.8754	0.9496
	Lasso	0.9236	0.0138	0.8966	0.9506
	Logistic regression	0.9282	0.0128	0.9032	0.9533
	Xgboost	0.9306	0.0128	0.9054	0.9557
	Ranger	0.9519	0.0103	0.9317	0.9720
	DiscreteSL	0.9519	0.0103	0.9317	0.9720
	SuperLearner	0.9644	0.0084	0.9480	0.9807
Dental	Mean of Y	0.5000	0.0493	0.4034	0.5966
	Decision tree	0.6537	0.0280	0.5989	0.7086
	Xgboost	0.6551	0.0270	0.6021	0.7081
	Ranger	0.7171	0.0250	0.6680	0.7662
	DiscreteSL	0.7213	0.0256	0.6711	0.7715
	Lasso	0.7412	0.0250	0.6921	0.7903
	Logistic regression	0.7417	0.0252	0.6924	0.7910
	SuperLearner	0.7421	0.0248	0.6935	0.7908
Combined craniodental	Mean of Y	0.5000	0.0493	0.4034	0.5966
	Decision tree	0.9060	0.0196	0.8675	0.9445
	Xgboost	0.9375	0.0116	0.9148	0.9602
	Logistic regression	0.9426	0.0111	0.9209	0.9643
	Lasso	0.9528	0.0104	0.9324	0.9731
	DiscreteSL	0.9528	0.0104	0.9324	0.9731
	Ranger	0.9549	0.0100	0.9353	0.9745
	SuperLearner	0.9722	0.0070	0.9585	0.9860

Table 6. Algorithm weight contributions to the SuperLearner ensembles.

Bony region	Algorithm	Mean (contribution to ensemble)	Standard deviation	Min	Max
Face	Ranger	0.4634	0.1058	0.2389	0.6044
	Logistic regression	0.4193	0.0373	0.3262	0.4779
	Xgboost	0.1159	0.0928	0.0000	0.3199
	Lasso	0.0013	0.0059	0.0000	0.0263
	Decision tree	0.0001	0.0004	0.0000	0.0017
	Mean of Y	0.0000	0.0000	0.0000	0.0000
Vault	Ranger	0.5004	0.1205	0.1910	0.7078
	Decision tree	0.3234	0.0935	0.1591	0.5442
	Logistic regression	0.1412	0.0520	0.0556	0.2234
	Xgboost	0.0350	0.0561	0.0000	0.1483
	Mean of Y	0.0000	0.0000	0.0000	0.0000
	Lasso	0.0000	0.0000	0.0000	0.0000
Base	Ranger	0.8878	0.0701	0.7068	0.9811
	Logistic regression	0.0758	0.0259	0.0189	0.1264
	Xgboost	0.0364	0.0590	0.0000	0.2168
	Mean of Y	0.0000	0.0000	0.0000	0.0000
	Lasso	0.0000	0.0000	0.0000	0.0000
	Decision tree	0.0000	0.0000	0.0000	0.0000
Crania	Ranger	0.4610	0.1162	0.2750	0.6789
	Logistic regression	0.1940	0.0859	0.0299	0.3193
	Lasso	0.1411	0.0753	0.0380	0.2882
	Decision tree	0.1267	0.1028	0.0000	0.3101
	Xgboost	0.0772	0.0826	0.0000	0.2452
	Mean of Y	0.0000	0.0000	0.0000	0.0000
Dental	Logistic regression	0.5591	0.0608	0.4472	0.6747
	Ranger	0.3582	0.0953	0.1797	0.5286
	Decision tree	0.0747	0.0719	0.0000	0.2339

	Xgboost	0.0080	0.0160	0.0000	0.0573
	Mean of Y	0.0000	0.0000	0.0000	0.0000
	Lasso	0.0000	0.0000	0.0000	0.0000
Combined craniodental	Lasso	0.4522	0.0918	0.2598	0.6602
	Ranger	0.1734	0.1048	0.0000	0.3853
	Xgboost	0.1700	0.0739	0.0416	0.2906
	Logistic regression	0.1319	0.0892	0.0000	0.3308
	Decision tree	0.0726	0.0755	0.0000	0.1891
	Mean of Y	0.0000	0.0000	0.0000	0.0000

Discussion

Performance of AUC analysis of the SuperLearner ensemble machine learning framework demonstrates strong potential of this methodology for sex estimation of archaeological remains. An important potential contribution of this research is that it reframes the problem of sex estimation as a predictive one and does not rely on the same assumptions of p-values, traditional hypothesis testing, or causal inference approaches. Additionally, the goal here was not to optimize any algorithms for maximum predictive accuracy, but to instead provide a gentle walkthrough of the process and to stimulate the reader into thinking about how this approach could be applied in their own research contexts. Instead, the focus was on model performance, standard errors, and confidence intervals. This method can also potentially be employed in the field to help resolve disagreements between experts or for indeterminate remains.

Results also support previous research that ensemble machine learning has strong potential for sex estimation in the bioarchaeological record (Muzzall et al., 2017). Although female-ness and male-ness were originally estimated using traits of the pelvis and skull in this study, results support previous research that indicates contrasts between male and female morphological and burial patterns in central Italy during the Iron Age (Coppa & Macchiarelli, 1982; Bondioli et al., 1986; Rubini, 1996; Muzzall & Coppa, 2019). Of particular interest was the general size differences between males and females despite their overlapping distributions. If the modeling process was strongly influenced by size however, it would be reasonable to expect that the dentition would have higher AUC values similar to that of the cranial data. Whether or not the antimeric substitution of left teeth for right teeth in the absence of a right-side tooth and/or the sheer amount of missingness influenced the much lower dental AUC is unknown. More cranial-dental comparisons are necessary to evaluate the reliability of the dentition in this framework. Among the three different cranial regions, the face had the highest AUC values, followed by the base and vault. This could provide further support of the utility of the face for population reconstruction despite its greater environmental plasticity compared to the base and vault due to sensory functions of sight, smell, and taste (Taubadel, 2009).

The ensembles themselves can be strengthened by including more algorithms and customizing them with varying hyperparameters (pre-training settings) to find the most accurate and best performing tunings (Bergstra & Bengio, 2012). Other considerations can be more thoroughly incorporated as well, such as different confusion matrix derivations to evaluate performance such as precision and recall to help highlight class imbalance problems, balanced estimator constructions, false discovery rate, and F1 score. Negative log-likelihood could also be used as the optimizer instead of nonnegative least squares. Other algorithms and methods might be more appropriate – only 6 algorithms with default settings were incorporated in this project but many others can be included in the ensemble (e.g., Bayesian additive regression trees, Chipman; George, & McCulloch, 2010). Features can be screened to identify more interpretable models and custom algorithms can be included to the researcher's exact specifications (see Kennedy, 2017 for the walkthrough in R). Moreover, deep learning – a subdiscipline of machine learning that utilizes multi-layered artificial neural networks for modeling, predicting, and understanding data – might be even more (Chollet & Allaire, 2017). When dataset sizes and the number of algorithms exceed personal compute potential, the software packages for analyses mentioned in this research have instructions to be run in parallel across multiple cores on a single computer or across multiple machines in cluster or remote settings. Perhaps of great interest to the bioarchaeologist, variable importance information can be extracted from the tree-based algorithms to see which cranial and dental dimensions have the highest weights for sex classification.

It is critical to note that due to the antiquity of the samples included in this research, the sexes of the individuals utilized were estimated macroscopically using features of the pelvis and skull and that the sexes were not actually certain thus making this study a sort of "estimation of estimations". Known-sex references samples should be a prerequisite for confirmation of methods presented here, and larger sample sizes might be important as well. This study is merely a demonstration of the methods and advertisement of the potential forgeneralized low rank imputation and ensemble machine learning processes in bioarchaeological and forensic contexts. Cadaver samples and skeletal collections would be particularly useful for testing procedures outlined here.

Ensemble machine learning techniques should be considered as part of the bioarchaeologist's toolkit as an additional method for comparison to macroscopic interrogations of the skeleton and dentition that we rely upon for reconstruction of the biological profiles of past humans. These techniques can potentially assist not only in bioarchaeological reconstructions, but also in forensic applications for identification of missing persons and perhaps even to material, faunal, and floral assemblages as well as mortuary studies and settlement organization. Furthermore, GLRMs warrant further exploration and should be considered by bioarchaeologists as a potentially strong data preprocessing tool when faced with missing data and analytical techniques that require full datasets for computation. Social scientists in general would benefit from updating their instrumentation with cross-validated ensemble machine learning techniques when research requires some variable(s) to be predicted.

Acknowledgements

I gratefully thank the UC Berkeley D-Lab, Alfredo Coppa, Chris J. Kennedy, and the Soprintendenza Archeologia d'Abruzzo, and the staffs from the Museo Antropologia de "Giuseppe Sergi" – Sapienza, Museo Paludi di Celano, Museo Archeologico Nazionale d'Abruzzo di Chieti, and Museo di Archeologico Nazionale di Campli. Aniket Kesari and Patrick Muzzall provided comments that improved the quality of this manuscript.

References

Bell, S., & Jantz, R. (2001). Neural network classification of skeletal remains. In G. Burenhult (Ed.) *Archaeological Informatics: Pushing The Envelope* (pp. 205-212). CAA2001. Oxford: Archaeopress.

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research 13,* 281-305.

Bondioli, L., Corruccini, R.S., & Macchiarelli, R. (1986). Familial segregation in the Iron Age community of Alfedena, Abruzzo, Italy, based on osteodental trait analysis. *American Journal of Physical Anthropology 71,* 393–400.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). Classification and regression trees. Belmont, California: Wadsworth.

Breiman, L. (2001a). Statistical modeling: The two cultures. *Statistical Science 16,* 199-231.

Breiman, L. 2001b. Random forests. *Machine Learning 45,* 5-32.

Buikstra, J.E., Frankenberg, S.R., Konigsberg, L.W. (1990). Skeletal biological distance studies in American physical anthropology: Recent trends. *American Journal of Physical Anthropology 82,* 1-7.

Buikstra, J.E., Ubelaker, D.H. (1994). Standards for data collection from human skeletal remains. Arkansas Archaeological Survey Research Series No. 44. Fayetteville, Arkansas: Arkansas Archaeological Survey.

Carson, E.A. (2006). Maximum likelihood estimation of human craniometric heritabilities. *American Journal of Physical Anthropology 131,* 169-180.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., Li, M., Xie, J., Lin, M., Geng, Y., & Li, Y. (2019). xgboost: Extreme Gradient Boosting. R package version 0.90.0.2. https://CRAN.R-project.org/package=xgboost

Chipman, H.A., George, E.I., & McCulloch, R.E. (2010). BART: Bayesian additive regression trees. *Institute of Mathematical Statistics – The Annals of Applied Statistics 1,* 266-298.

Chollet, F., & Allaire, J.J. (2017). Deep learning with R. New York: Manning.

Coppa, A., & Macchiarelli, R. (1982). The maxillary dentition of the Iron-Age population of Alfedena (Middle-Adriatic Area, Italy). *Journal of Human Evolution 11,* 219–235.

Coppa, A., Cucina, A., Lucci, M., Mancinelli, D., & Vargiu, R. (2007). Origins and spread of Agriculture in Italy: A nonmetric dental analysis. *American Journal of Physical Anthropology 133,* 918-930.

Cunningham SJ. (1997). Machine learning applications in anthropology: Automated discovery over kinship structures. *Computers and the Humanities 30,* 401-406.

Czibula, G., Ionescu, V.S., Miholca, D.L., & Mircea, I.G. (2016). Machine learning-based approaches for predicting stature from archaeological skeletal remains using long bone lengths. *Journal of Archaeological Science 69,* 85-99.

Devor, E.J. (1987). Transmission of human cranial dimensions. *Journal of Craniofacial Genetics and Developmental Biology 7,* 95- 106.

Dietterich, T.G. (2000). Ensemble methods in machine learning. In G. Goos, J. Hartmanis, & J. van Leeuwen (Eds.). *Lecture Notes in Computer Science 1857,* 1-15.

Dobson, A.J. (1990). *An Introduction to Generalized Linear Models.* London: Chapman and Hall.

Efron, B., &Gong, G. (1982). A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician 37,* 36-48

Freund, Y., & Schapire R.E. (1999). A short introduction to boosting. *Journal of Japanese Society for Artificial Intelligence 14,* 1-14.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software 33*, 1-22.

Fu, Q., Rudan, P., Pääbo, S., & Krause, J. (2012). Complete mitochondrial genomes reveal Neolithic expansion into Europe. *PLoS ONE 7(3),* e32473.

Garvin, H.M., &Ruff, C.B. (2012). Sexual dimorphism in skeletal browridge and chin morphologies determined using a new quantitative method. *American Jounral of Physical Anthropology 147,* 661-670.

Ghirotto, S., Tassi, F., Fumagalli, E., Colonna, V., Sandionigi, A., Lari, M., Vai, S., Petiti, E., Corti, G., Rizzi, E., De Bellis, G., Caramelli, D., & Barbujani, G. (2013). Origins and evolution of the Etruscans' mtDNA. *PLoS ONE 8(2),* e55519.

Hanley, J.A., & McNeil, B.J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology 143,* 29-36.

Hefner, J.T., & Ousley, S.D. (2014). Statistical classification methods for estimating ancestry using morphoscopic traits. *Journal of Forensic Sciences 59,* 883-890.

Herrera, B., Hanihara, T., & Godde, K. (2014). Comparability of multiple data types from the Bering Strait region: Cranial and dental metrics and nonmetrics, mtDNA, and Y-Chromosome DNA. *American Jounral of Physical Anthropology 54,* 334-348.

Hillson, S., FitzGerald, C., & Flinn, H. (2006). Alternative dental measurements: Proposals and relationships with other measurements. *American Jounral of Physical Anthropology 126,* 413-426.

Ionescu, V.S., Teletin, M., & Voiculescu, E.M. (2016). Machine learning techniques for age at death estimation from long bone lengths. In *2016 IEEE 11th International Symposium on Applied Computational Intelligence and Informatics (SACI)* pp. 457-462. Timisoara, Romania.

Ionescu VS, Czibula G, & Teletin M. 2018. Supervised learning techniques for body mass estimation in bioarchaeology. In: V. Balas, L. Jain, & M. Balas (Eds.) *Soft Computing Applications*. SOFA 2016. Advances in Intelligent Systems and Computing 634. Springer.

Irurita Olivares, J., & Alemán Aguilera, I. (2016). Validation of the sex estimation method elaborated by Schutkowski in the Granada Osteological Collection of identified infant and young children: Analysis of the controversy between the different ways of analyzing and interpreting the results. *International Journal of Legal Medicine 130,* 1623-1632.

Jackes, M. (2011). Representativeness and bias in archaeological skeletal samples. In: S.C. Agarwal, & B.A. Glencross (Eds.) *Social Bioarchaeology* (pp. 107-145). West Sussex, UK: Wiley-Blackwell.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An intro to statistical learning: With applications in R. New York: Springer.

Kennedy, C. (2017). Guide to SuperLearner. https://cran.r-project.org/web/packages/SuperLearner/vignettes/Guide-to-SuperLearner.html

Kennedy, C. (2020). ck37r: Chris Kennedy's R toolkit. R package version 1.0.3 https://github.com/ck37/ck37r

Kenyhercz, M.W., & Passalacqua, N.V. (2016). Missing data imputation methods and their performance with biodistance analyses. In M.A. Pilloud, & J.T. Hefner (Eds.) *Biological Distance Analysis – Forensic and Bioarchaeological Perspectives* (pp. 181-194). San Diego: Academic Press.

Konigsberg, L.W., Algee-Hewitt, B.F.B., & Steadman, D.W. (2009). Estimation and evidence in forensic anthropology: Sex and race. *American Journal of Physical Anthropology 139,* 77-90.

Krishan, K., Chatterjee, P.M., Kanchan, T., Kaur, S., Baryah, N., & Singh, R.K. (2016). A review of sex estimation techniques during examination of skeletal remains in forensic anthropology casework. *Forensic Science International 261,* 165.e1-165.e8.

Lantz, B. (2015). Machine learning with R: Discover how to build machine learning algorithms, prepare data, and dig deep into data prediction techniques with R, 2nd ed. Birmingham, UK: Packt Publishing.

Martínez-Abadías, N., Esparza, M., Sjøvold, T., González-José, R., Santos, M., & Hernández, M. (2009). Heritability of human cranial dimensions: comparing the evolvability of different cranial regions. *Journal of Anatomy 214,* 19-35.

Miholca, D.L., Czibula, G., Mircea, I.G., & Czibula, I.G. (2016). Machine Learning Based Approaches for Sex Identification in Bioarchaeology. In *2016 18th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC)* pp. 311-314. Timisoara, Romania.

Muttoni, G., Scardia, G., Kent, D.V., Swisher, C.C., & Manzi, G. (2009). Pleistocene magnetochronology of early hominin sites at Ceprano and Fontana Ranuccio, Italy. *Earth and Planetary Science Letters 286,* 255-268.

Muzzall, E., Kennedy, C.J., & Culich, A. (2017). Ensemble machine learning for sex prediction of a worldwide craniometric dataset. Poster presented at the Berkeley Institute for Data Science Spring 2017 Data Science Faire. https://github.com/EastBayEv/Ensemble-machine-learning-for-sex-prediction-of-a-worldwide-craniometric-dataset/blob/master/Ensemble%20machine%20learning%20for%20sex%20prediction%20of%20a%20worldwide%20craniometric%20dataset.pdf

Muzzall, E., & Coppa, A. (2019). Temporal and spatial biological kinship variation at Campovalano and Alfedena in Iron Age Central Italy. In: C. Tica, & D.L. Martin (Eds.) *Bioarcheology of Frontiers and Borderlands* (pp. 107-132). University Press of Florida.

Ortiz, A.G., Costa, C., Silva, R.H.A., Biazevic, M.G.H., & Michel-Crosato, E. (2020). Sex estimation: Anatomical references on panoramic radiographs using machine learning. *Forensic Imaging* 20:200356.

Pink, C.M. (2016). Forensic ancestry assessment using cranial nonmetric traits traditionally applied to biological distance studies. In M.A. Pilloud, & J.T. Hefner (Eds.) *Biological Distance Analysis – Forensic and Bioarchaeological Perspectives* (pp. 213-230). San Diego: Academic Press.

Polley, E.C., & van der Laan, M.J. (2010). Super Learner in prediction. *UC Berkeley Division of Biostatistics Working Paper Series Paper 266,* 1-19.

Polley, E., LeDell, E., Kennedy, C., & van der Laan, M. (2019). SuperLearner: Super Learner Prediction. R package version 2.0-26 https://CRAN.R-project.org/package=SuperLearner

Porto, F.P., Correia Lima, L.N., Pihneiro Flores, M.R., Valsecchi, A., Ibanez, O., Machado Palhares, C.E., & de Barros Vidal, F. (2019). Automatic cephalometric landmarks detection on frontal faces: An approach based on supervised learning techniques. *Digital Investigation 30,* 108-116.

Roseman, C.C. (2004). Detecting interregionally diversifying natural selection on modern human cranial form by using matched molecular and morphometric data. *Proceedings of the National Academy of Sciences 101,* 12824-12829.

Roseman, C.C., & Weaver, T.D. (2004). Multivariate apportionment of global human craniometric diversity. *American Jounral of Physical Anthropology 125,* 257-263.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Pyschological Review 65,* 386-408.

Rubini, M. (1996). Biological homogeneity and familial segregation in the Iron Age population of Alfedena (Abruzzo, Italy), based on cranial discrete traits analysis. *International Journal of Osteoarchaeology 6,* 454–462.

Samuel, A.L. (1959). Some studies in machine learning using the game of checkers. *IBM Jounral of Research & Development 3,* 207-226.

Scozzari, R., Cruciani, F., Pangrazio, A., Santolamazza, P., Vona, G., Moral, P., Latini, V., Varesi, L., Memmi, M.M., Romano, V., De Leo, G., Gennarelli, M., Jaruzelska, J., Villems, R., Parik, J., Macaulay, V., & Torroni, A. (2001). Human Y-chromosome variation in the western Mediterranean area: Implications for the peopling of the region. *Human Immunology 62,* 871-884.

Sierp, I., & Henneberg, M. (2015). The difficulty of sexing sekeltons from unknown populations. *Jounral of Anthropology* 908535.

Sjøvold, T. (1984). A report on the heritability of some cranial measurements and non-metric traits. In G.H. Van Vark, & W.W. Howells (Eds.) *Multivariate Statistical Methods in Physical Anthropology* (pp. 223-246). Dordrecht: Reidel Publishing Company.

Slemenda, C.W., Reister, T.K., Hui, S.L., Miller, J.Z., Christian, J.C., & Johnston, C.C. (1994). Influences on skeletal mineralization in children and adolescents: Evidence for varying effects of sexual maturation and physical activity. *Journal of Pediatrics 125,* 201-207.

Strauss, A., & Hubbe, M. (2010). Craniometric similarities within and between human populations in comparison with neutral genetic markers. *Human Biology 82,* 315-330.

Sutter, R.C. (2003). Nonmetric subadult skeletal sexing traings: I. A blind test of the accuracy of eight previously proposed methods using prehistoric known-sex mummies from northern Chile. *Journal of Forensic Sciences 48,* 927-935.

Taubadel, N.V.C. (2009). Revisiting the Homoiology Hypothesis: The Impact of Phenotypic Plasticity on the Reconstruction of Human Population History from Craniometric Data. *Journal of Human Evolution 57,* 179-190.

Turing, A.M. (1950). Computing machinery and intelligence. *Mind 59,* 433-460.

Udell, M., Horn, C., Zadeh, R. & Boyd, S. (2016) Generalized low rank models. *Foundations and Trends in Machine Learning 9, 1-118.*

van der Laan, M.J., Polley, E.C., & Hubbard, A.E. (2007). Super Learner. *Statistical Applications in Genetics and Molecular Biology 6,* Article 25 pp. 1-21.

Wang, Y. (2002). Is obesity associated with early sexual maturation? A comparison of the association in American boys versus girls. *Pediatrics 110,* 903-910.

Welling, M. (2015). Are ML and statistics complimentary? Roundtable discussion at the 6th IMS-ISBA meeting on “Data Science in the Next 50 Years”.

Weiss, K.M. (1972). On the systematic bias in skeletal sexing. *American Jounral of Physical Anthropology 37,* 239-249.

Wickham, H. (2016). ggplot2: Elegant graphics for data analysis. New York: Springer-Verlag. https://ggplot2.tidyverse.org

Witherspoon, D.J., Wooding, S., Rogers, A.R., Marchani, E.E., Watkins, W.S., Batzer, M.A., & Jorde, L.B. (2007). Genetic similarities within and between human populations. *Genetics 176,* 351-359.

Wright, N., & Ziegler, A. (2017). ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software 77,* 1-17.