UNIVERSITY OF CAPE TOWN

Ancestry-Independent Osteometric Sex Estimation from Selected Postcranial Skeletal Elements of South Africans: A Machine Learning Approach

Author:

Michelle Scott

Supervisor:

Associate Professor Jacqui Friedling

Dissertation presented for the degree of

Master of Science in Medicine (MSc (Med)) in Biological Anthropology

in the Department of Human Biology

2022/02/13

DECLARATION

I, Michelle Ashleigh Scott, hereby declare that the work on which this dissertation/thesis is based is my original work (except where acknowledgements indicate otherwise) and that neither the whole work nor any part of it has been, is being, or is to be submitted for another degree in this or any other university.

I empower the university to reproduce for the purpose of research either the whole or any portion of the contents in any manner whatsoever.

Signature:

Signed by candidate

Date: 2022/02/13

D19 PLAGIARISM DECLARATION

This thesis/dissertation has been submitted to the Turnitin module (or equivalent similarity and originality checking software) and I confirm that my supervisor has seen my report and any concerns revealed by such have been resolved with my supervisor.

Name: Michelle Ashleigh Scott

Student Number: sctmic013

Signature: Signed by candidate

Date: 2022/02/13

ACKNOWLEDGEMENTS

This is the bit where I get to write from the heart and not be professional. I am unbelievably lucky to have been surrounded by a village of loving and supportive people throughout my life who have shown a keen interest in my studies and have encouraged me to pursue my passions. There are a number of people who require my particular gratitude. My supervisor, Jacqui Friedling, your guidance has been greatly appreciated. To the very helpful and accommodating Dr Billings at WITS and the Raymond A. Dart Collection, thank you for hosting me and allowing me access to your repository. Similarly, thank you to the UCT Skeletal Repository for granting me access and Caroline Powrie for assisting in my data collection journey. Great Uncle Robin, thank you so much for opening up your home to me and making my dreaded data collection trip to scary Johannesburg the highlight of my year. I'd like to thank Paul Fatti (Emeritus Prof. of Statistics at WITS) for his food for thought advice on how to proceed when I found myself at a roadblock with normality testing. My wonderful, crazy family – you all drive me mad daily, but I am so immensely grateful for all of the things that you have done for me. I feel incredibly lucky and privileged to be so blessed. Thank you all for listening to my rants and putting up with my stress-induced foul moods. Granny, you're right, most problems are much easier solved after a cookie and a cuddle. Granddad, thank you for curating my love of science from a young age with trips to the science centre, explaining how rock formations are formed, pressurised coke bottle cars and our many travels around South Africa. I am eternally grateful to have had your influence in my life. My mom, thank you for teaching me independence, accountability, and work-ethic. Uncle Ian, I miss you every day and I will always have a million things that I wish I could say to you. Thank you for filling my life with joy, love, and generosity. David, it is really quite impossible to thank you enough. Thank you for making me approximately 1 million cups of tea, writing the Bone Collector for me, confiscating my phone, pushing me to try things outside of my academic comfort zone, teaching me how to write my first FOR loop, telling me off for following bad coding practices, troubleshooting my broken code, editing my thesis, and keeping me happy and bouncy.

For my uncle, Ian Scott,

who inspired a love of knowledge

and is sorely missed.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Parameters:

AD	Acetabulum Diameter
BB	Bicondylar Breadth of Femur
EB	Humerus Epicondylar Breadth
FHD	Femoral Head Diameter
GB	Glenoid Breadth
GL	Glenoid Length
HHD	Vertical Head Diameter of Humerus
LCL	Lateral Condylar Length of Femur
MCL	Medical Condylar Length of Femur
TAD	Transverse Acetabulum Diameter
THD	Transverse Femoral Head Diameter
TPB	Tibia Proximal Breadth

General:

CT	Computed Tomography
Dart	Raymond A. Dart Human Skeletal Collection
DNA	Deoxyribonucleic Acid
ICT	Infancy-Childhood Transition
PCR	Polymerase Chain Reaction
SA	South Africa(n)
SAAA	South African of African Ancestry
SAEA	South African of European Ancestry
SAMA	South African of Mixed Ancestry

(note: ‘Mixed’ refers to an ancestry group whilst ‘mixed’ does not)

SUN	Stellenbosch University
UCT	University of Cape Town
UP	University of Pretoria
WITS	University of the Witwatersrand

Statistical:

2/3D	Two/Three Dimensional
CCC	Lin's concordance correlation coefficient
CI	Confidence interval
CV	Cross-validation
DFA	Discriminant function analysis
Discr	Discriminant Function Analysis
Ho	Holdout
KNN	K-nearest neighbour
LDA	Linear discriminant analysis
LR	Logistic regression
max	Maximum
ML	Machine Learning
min	Minimum
n	Number of individuals
NB	Naïve Bayes
Q-Q	Quantile-Quantile
r	Pearson's r-statistic
REP	Reduced Error Pruning
SD	Standard deviation
SE	Standard Error
SEM	Standard Error of Measurement
Tree	Binary decision tree

ABSTRACT

Sex estimation, as part of a biological profile, has the power to halve the number of possible identities of unidentified skeletal remains. Postcranial elements have been studied in South Africa (SA) for the purpose of sex estimation and have often proven to be more accurate than the cranium. Estimation techniques using postcranial elements in SA almost exclusively utilise discriminant analysis to evaluate sex, but international publications have shown success using alternative machine learning (ML) algorithms. SA methods and standards are often restricted by limited sample size, lack of robust statistical techniques in older publications and, the prerequisite of known or estimated ancestry. Most methods are specific to SA African, European or, more recently, Mixed ancestry groups and are unreliable when ancestry is unknown. The aim of this study was to apply a series of ML algorithms to train ancestry-independent sex classification models using postcranial osteometric measurements from the cadaveric skeletal remains of modern South Africans, focussing on long bone joints. The study consisted of a roughly demographically representative, pooled sample, of 650 South Africans (325 male, 325 female). 12 osteometric measurements were taken from available left- and, or right-sided bones for each individual. All 12 mensurations were sexually dimorphic and differences between left- and right-sided bones were negligible. The dataset was subjected to ML algorithm training using univariate and multivariate predictor combinations to train decision tree, ensemble, *k*-nearest neighbour, Naïve Bayes, and discriminant function models. The best performing ML algorithm, given the sample size and available predictors was discriminant function analysis. Univariate model accuracies ranged from 80.5-89.1% and multivariate model accuracies ranged from 84.5%, using 2 predictors, to 92.8%, using 12 predictors. An optimised 3-predictor model was able to predict sex with 92.7% accuracy. Results from this study were comparable to those using ancestry-specific models and non-ancestry-specific models, where available. Findings from this study suggested that the inclusion of ancestry, when predicting sex using the elements examined, is not necessary. Existing ancestry-dependent models were unable to outperform new ancestry-independent models especially when considering the added error associated with evaluating ancestry before sex could be assessed.

CHAPTER 1: INTRODUCTION

Hikers stumble upon human skeletal remains. A mass grave is discovered during excavation for a new skyscraper. The burnt and dismembered remains of victims from a violent gang battle are uncovered in a clandestine grave. A farm dog arrives home from the fields carrying a human femur. Fragments of human bone wash up on the shore. These illustrate just a few of the kinds of cases within the purview of the forensic anthropologist (Christensen & Crowder, 2009). In all of these instances, the construction of a biological profile is imperative to assist in the correct identification of the human skeletal remains (Patriquin, Steyn & Loth, 2005).

Accurate identification of an individual plays a vital role in medicolegal investigations ranging from individual homicide cases to mass deaths resulting from massacres and natural disasters (Franklin, 2010). As a result of the high crime rate in South Africa, the accurate estimation of biological identifiers is important to reduce backlog and ensure both social and criminal justice for victims (Seedat, Niekerk, et al., 2009). A biological profile currently consists of an assessment of age-at-death, ancestry, biological sex and living stature (Lundy, 1998). These variables help to refine the possibilities of identification. Sex, as a biological variable, is incredibly valuable as it has the power to halve the number of possible identities of the remains and is essential for the reconstruction of age-at-death and living stature (Loth & Işcan, 2000).

While sex assessments are rarely executed on individuals preceding 12 years of age, as human males and females are largely androgynous until pubescence (Patriquin, Steyn & Loth, 2005; Christensen & Crowder, 2009), they can be of much value in the correct identification of adult skeletal remains. This is because, although age-at-pubescence varies by individual and population, skeletal changes are predictable as a result of degeneration and pathologies associated with advancing age and lifestyle factors over the course of any individual’s lifetime (Vance, 2007).

During adult life, sex is a discrete trait determined by genetic makeup and can be readily discerned by primary sexual characteristics, such as external genitalia. These easily recognisable traits allow persons to be classified into one of two categories – male or female. However, in the human skeleton, as opposed to the human body, the loss of soft tissue makes identification more complex as one must rely on secondary sexual characteristics (features which appear during puberty) which may be largely influenced by environmental factors (Nikitovic, 2018). This leads one to consider the subject of sexual dimorphism.

Sexual dimorphism, the size and shape-based differences associated with secondary sexual characteristics of males and females within a species, exists on a continuum in the human skeleton, with a large degree of overlap between the sexes (Nikitovic, 2018; Jerković et al., 2020).

A growth spurt and an escalation in muscular tissue, which is particularly noticeable in males, strikes at puberty (Wells, 2007). These skeletal and soft tissue modifications are exacted by several interrelated factors including intrinsic heredity and endocrine functions, and extrinsic factors like nutrition (Bogin, 1995). The inhibitory effect of oestrogen on female periosteal apposition, in conjunction with the prominent growth spurt in males, contribute to adult females generally having lower bone mass and smaller size than their male counterparts (Stulp & Barrett, 2016). Complete epiphyseal union, marking the end of the growth phase, is observed as late as age 34 in some bones (Owings & Myers, 2005). Bony features, such as the ventral arc of the pubis, are only easily distinguished around age 23 before succumbing to age-related degeneration and arthritis and once again becoming indistinguishable (Sutherland & Suchey, 1991).

Difficulties in determining sex in the human skeleton are confounded by the fact that remains are often fragmentary (Kelley, 1979), populations vary in their expression of certain features (Kotěrová et al., 2017) and the identification of specific traits may be dependent on the experience of an observer (Walrath, Turner & Bruzek, 2004).

Sex estimation in South Africa is generally performed using molecular, morphological, geometric morphometric and metric methods.

Molecular sex estimation methods analyse DNA extracted from hard tissue and evaluate sex by isolating and amplifying genetic information contained in the sex chromosomes (Latham & Miller, 2019). Molecular sex estimation methods have many merits, including: their high level of accuracy; ability to estimate sex regardless of fragmentation or missing skeletal elements; and, importantly, their ability to determine sex even when remains belong to a minor or individual of unknown ancestry, nationality, or chronologic age (Krishan et al., 2016). Despite their many merits, molecular analytics are costly, specialized, and time-consuming. For these reasons, molecular methods are currently not easily accessible or user-friendly and cannot be applied quickly in the field for archaeological or forensic cases.

Morphological methods consist of visual observation of bony landmarks displaying sexual dimorphism. These methods are quick to apply, are distinguishable despite variations in populations across time, and do not require any specialist equipment, thus making them easily applicable in the field. However, visual observation of shape-dependent traits is highly dependent on subjectivity, and the expertise of the observer (Walrath, Turner & Bruzek, 2004). Their accuracy in separating the sexes is now being challenged and reassessed using modern morphometric techniques (Steyn, Pretorius & Hutten, 2004). The subjectivity associated with morphological methods make them suboptimal as evidentiary matter in medico-legal proceedings.

Geometric morphometrics observe and quantify shape *and* size differences in bones. This utilises the power of shape analysis whilst eliminating the subjectivity of any size information associated with it. Shape and size can be analysed in both 2 and 3 dimensions using digitising equipment or scans. Problems arise when using these techniques in the field as precise positioning of bones and camera equipment is usually required. Software is also often needed to perform the time-consuming process of assigning landmarks and co-ordinates (Christensen & Crowder, 2009). Thus, whilst subjectivity is reduced, similarly to molecular methods, geometric morphometrics are often overlooked due to their perceived complexity.

Metric methods involve single or multiple measurements which are substituted into mathematical functions to estimate sex. Most metric sex classifiers in South Africa (SA) rely on discriminant function analysis, although little justification exists for why other statistical methods have not been employed to tackle the sex classification problem (Bidmos, Gibbon & Štrkalj, 2010). Within the sphere of metric sex estimation outside of SA, machine learning (ML) classifiers, in addition to discriminant function analysis, have recently proven to be useful (Navega et al., 2015; Curate, Umbelino, et al., 2017; Coelho & Curate, 2019).

Not all parts of the skeleton show reliable sexually dimorphic morphology. In these instances - when differences are largely size-based, and when incomplete or damaged remains are found - the observer may need to utilise less dimorphic skeletal elements. In these cases, metric methods are applied. Metrics are advantageously easy to apply in the field. Their accuracy is quantifiable, observer error and repeatability are easily assessed, and results are simple to interpret regardless of experience. These features make metric methods the ideal candidate for sex classification, but no methods are infallible. Problems arise with overlap between sexes and the elevated specificity of discriminant functions to a particular population. As a result, most SA metric methods require the predetermination of ancestry to select the correct mathematical discriminant functions to estimate sex (Steyn & Işcan, 1997; Asala, 2001).

The topic of ancestry is covered in Chapter 2. Ancestry grouping is fraught with issues including the vast generalisations needed to segregate a highly admixed, diverse population into the 3 socially and bureaucratically defined groups featured in most SA metric sex classification methods (Işcan & Steyn, 2013; Petersen et al., 2013; Cunha & Ubelaker, 2020). Genetic comparisons between ancestral groups have shown that around 85% of all human variation occurs within groups with only a small percentage occurring between groups (Gannett, 2014) and evidence has been provided to suggest that the size-based variation observed between SA ancestry groups (Arendse, 2018) may be a consequence of the socio-political landscape over the last century rather than intrinsic biological differences (Henneberg & van den Berg, 1990). Multivariate postcranial ancestry estimation using Fordisc 3.1 provided accuracies of only 79.0% (Liebenberg et al., 2019) meaning that compounding error from incorrect ancestry estimation could drastically lower the effective accuracy of any sex estimation method which requires the predetermination of ancestry (Bidmos & Dayal, 2004).

AIMS AND OBJECTIVES

The aim of this research is to use generally well-preserved, sexually dimorphic postcranial osteometric parameters - measured from the glenoid fossa, head of the humerus, distal humerus, acetabulum, head of the femur, distal femur, and proximal tibia of skeletonised South African cadaveric individuals - to develop new metric sex classifiers which can be applied to skeletal remains of unknown ancestry. This will be achieved by completing the following objectives:

1. Measure parameters on South African individuals from documented South African cadaveric skeletal collections.
2. Understand the structure of the sample and resulting dataset, and investigate potential confounding factors by assessing:
    a. Errors and underlying distributions
    b. Bilateral variation
    c. Sexual Dimorphism
    d. Effects of advanced age
3. Develop sex estimation methods using Machine Learning (ML) techniques, which can be applied to skeletal remains for which ancestry is unknown.
4. Compare the prediction accuracies of new methods to those achieved using traditional discriminant functions and sectioning points.

RESEARCH QUESTION

Can ML techniques be used to derive ancestry-independent metric sex estimation methods which can correctly estimate sex more accurately than traditionally derived, ancestry-dependent methods?

HYPOTHESIS

Ancestry-independent sex classification methods can predict sex with accuracies equal to ancestry-dependent sex classification methods.

CHAPTER 2: BACKGROUND

In the South African medico-legal arena, the biological anthropologist's job is to refine the possibilities of who unidentified human remains may have been in life by constructing a biological profile (Patriquin, Steyn & Loth, 2005). Comparisons to determine age-at-death, sex and stature all rely on inherent skeletal variation between individuals.

2.1 SKELETAL VARIATION

Many theories exist to explain the differences in body size between and within modern population groups. Causational factors in adult stature include genetics (Weedon & Frayling, 2008; Petersen et al., 2013; German et al., 2015; Rawlik, Canela-Xandri & Tenesa, 2016; Stulp & Barrett, 2016) as well as a number of external features including: ecogeographic influences (Cowgill et al., 2012; Wells, 2012), developmental nutrition (Akachi & Canning, 2007; Victora et al., 2008), health and disease (Deaton, 2007; Limony, Friger & Hochberg, 2013), and economics (Inwood & Masakure, 2013; Stulp & Barrett, 2016). Additional proposed contributors to diversity in South African population groups comprise ancestral input as well as historical socio-politics, and subsequent economic conditions (Isaacs-Martin & Petrus, 2012; Inwood & Masakure, 2013; Petersen et al., 2013; Montinaro et al., 2017).

Normal human height variation is genetically influenced by between 44 and 200 loci in human DNA (Weedon & Frayling, 2008; German et al., 2015). Whilst genetics may contribute to normal human height, Europeans are now 13cm taller than they were 150 years ago (German et al., 2015). It is unlikely that these changes are the result of 6 generations worth of changes in gene sequence alone, with other contributing factors being a far more likely explanation.

Age-at-transition from infancy to childhood (ICT), marked by a growth spurt associated with growth hormone, has been shown to negatively correlate with final adult height (Hochberg & Albertsson-Wikland, 2008). A delay in ICT is an adaptive evolutionary strategy in response to an energy crisis during infancy and usually leads to shorter adult stature (Hochberg et al., 2011).

Nutritional deprivation during infancy, fleeting as it may be, is thus capable of eliciting a lifelong influence on gene expression via epigenetic mechanisms and is capable of altering both chromatic conformation and the accessibility of transcription factors involved in hormonal regulation (Hochberg & Albertsson-Wikland, 2008; Victora et al., 2008). ICT alone is shown to have contributed up to 50.0% of secular trend over the last 150 years and is responsible for 28.0% of human height variation (German et al., 2015).

It is therefore far more plausible that ICT, rather than changes in gene sequence alone, has contributed to Europeans being so much taller now than they were a century and a half ago. This can likely be attributed to a substantial improvement in food security over the aforementioned time period (German et al., 2015).

Another contributor to skeletal variation is diet, which plays a vital role in human growth. An 'energy crisis' during infancy is often the result of malnutrition and has lifelong effects on health with the capacity to influence future generations (Victora et al., 2008). Variation in height can be correlated with protein and general caloric intake at birth and during adolescence with shorter stature recorded when nutritional intake was insufficient (Akachi & Canning, 2007). Often, in cases where the household income is low, there is a disparity between male and female nutrition leading to a lower rate of stunting in males (Rawlik, Canela-Xandri & Tenesa, 2016). There is some evidence from developed countries that up to 80.0% of height variation reflects genetic variation (Silventoinen et al., 2003) but this is less influential in developing countries where poverty, and the resulting malnutrition, plays a major role in height variation (Akachi & Canning, 2007).

The impact of low per capita household income on skeletal variation is not limited to diet alone. Low income also results in limited access to medical care, as well as exposure to pathogens during development, which both have a major knock-on effect on health, growth, and ultimately adult stature (Deaton, 2007). Disease stunts growth and leads to shorter adult height, whilst poverty leads to malnutrition which, in turn, causes reduced adult stature (Inwood & Masakure, 2013; Limony, Friger & Hochberg, 2013).

A human's capacity to reach their full growth potential is therefore reliant on not only genetics but also the environment in which they grow up (Hawley et al., 2009). Temporal trends towards increasing adult stature and body size have been correlated with the removal of growth constraints via improved nutritional provision, health care and socioeconomic environments (Bogin, 1995). Globally, there is a trend towards increasing height as child undernutrition is reduced (Cole, 2000; Victora et al., 2008). In South Africa, positive secular trends of increasing mean height were observed between 1880 and 1990 in all SA ancestry groups. These trends were very weak (0.5mm average per decade) and much less pronounced than trends seen in non-SA European samples (Henneberg & van den Berg, 1990). Hawley et al. (2009) notes that there has been a significant change in environment for South Africans over the last 150 years. Improved environment, access to healthcare and better childhood nutrition are all attributed to the rise in life expectancy from 50 years in 1960 to 62 years in 1990 (Hawley et al., 2009). Projections suggest that mean life expectancy for South Africans will reach 70 years by 2030 (Statistics South Africa, 2018). Fears that secular changes may lead to increased difficulty in sex estimation have been disproven with evidence showing that secular trend has led to increased sexual dimorphism and improved sex prediction (Klales, 2016).

Biogeography Is thought to play a role in variation between populations. According to Petersen et al. (2013), "within-population genetic diversity is greatest within Africa" and "between-population genetic diversity is directly proportional to geographic distance between groups" (Petersen et al., 2013). Climatic data has been shown to correlate with stature, BMI, bi-iliac breadth, and weight (Cowgill et al., 2012). Individuals born and raised at high altitudes are usually short in stature with high BMIs relative to height as well as wider bodies and a shorter limb-to-trunk ratio. This is compared to those who develop at lower altitudes and have lower BMIs, longer limbs, and narrower bodies (Klaus, 2014). These factors lead to regional diversity in skeletal size and shape.

Due to the contributing factors to skeletal variation outlined above, prior knowledge of the nationality and/or ancestry of the individual to be sexed is necessary to avoid applying discriminant functions derived for a well-known population or group to remains from a different group, time, or geographical region (Dillon, 2014; Introna et al., 2015; Kotěrová et al., 2017). Generating standards which can be applied to local populations from specific geographical regions (i.e., South Africa) and updating these standards to accommodate temporal variations and secular trend is vital (Steyn & Işcan, 1997).

2.2 ANCESTRY

The use of race, ethnicity, ancestry, and biogeographic ancestry in the biological sciences is widely debated (Shields et al., 2005; Gannett, 2014). 'Race' terminology in biological anthropology was mostly replaced by 'ancestry' in the early 2000s (Ross & Pilloud, 2021). Biogeographical ancestry, which is defined by Shields et al. (2005) as "the component of ethnicity that is biologically determined and can be estimated using genetic markers that have distinctive allele frequencies for the populations in question", is intended to replace socio-politically constructed racial categories which are not anthropologically or scientifically based with ones which are supposedly rooted in biological differences (Shields et al., 2005). Gannett (2014), however, punctuates fears that any separation of demographic groups may well risk suggesting that "there are fundamental biological and behavioural differences between racial groups" and emphasises that ancestral groupings are ineffective in a modern world where there is a large degree of admixture (Gannett, 2014). More recently, Ross & Pilloud (2021) have recommended a new term, population affinity, which is intended to be a statistical approach to describing the underlying population structure based on microevolutionary forces like historical events and how they have shaped modern human variation.

Within South Africa, ancestry, in a biological sense, is constructed based on skeletal variation associated with broad biogeographical ancestral categories related to geographic origin and historical peopling i.e., African, Asian, or European (Işcan & Steyn, 2013). This empirically deduced judgement is however, not in line with the socially and medicolegally recognized terminology in this country. People in South Africa (SA) currently self-classify as and identify with population aligned social racial categories which are also used by government as legal terms (Tawha et al., 2020). These groups are generally referred to as "Black', "Coloured' and "White" and are not exclusively associated with general geographical origin (Petersen et al., 2013). This racial terminology was also used by both the UCT Cadaveric Repository and the University of Stellenbosch Kirsten Collection but is slightly altered at the University of the

Witwatersrand (Wits) Raymond A. Dart collection, which has replaced “Black” with “African” (Maass & Friedling, 2019; Bidmos et al., 2009; Alblas & Greyling, 2018). UCT has recently begun a campaign to eliminate race and ancestry, but these updates have yet to be published.

The SA population is unusually diverse, as SA was historically occupied by Khoesān hunter-gatherers until roughly 2000 years ago when archaeological signatures of domesticated plants and animals, iron-work and sedentary settlement signal the arrival of immigrant Bantu-speaking farmers (Ribot et al., 2010). For the purpose of this study, the ancestral groups within SA are broadly classified as South Africans of African Ancestry (SAAA), South Africans of European Ancestry (SAEA) and South Africans of Mixed Ancestry (SAMA).

SAAA, bureaucratically recognized as ‘Black’ and descended from Bantu-speaking peoples of East and West Africa, which migrated south between 300 and 1500 years ago (Henn et al., 2008; Lambert & Tishkoff, 2009; Petersen et al., 2013; Montinaro et al., 2017; Skoglund et al., 2017) are the largest ancestry group in SA, making up 80.7% of the population (Statistics South Africa, 2019). The bureaucratically recognized ‘White’ group, hereafter described as SAEA, are predominantly descendants of immigrants and colonial settlers from England, France, Germany, Greece, Italy, Portugal, and the Netherlands (De Wit et al., 2010; Işcan & Steyn, 2013; Liebenberg, L’Abbé & Stull, 2015; Montinaro et al., 2017). This group, which constitutes 7.9% of the SA population (Statistics South Africa, 2019), is considered to be the most homogenous in South Africa because of the founder’s effect, ecogeographic differences and limited admixture with other established groups as a result of the South African socio-political landscape over the last 350 years (Petersen et al., 2013). Finally, SAMA, the group bureaucratically defined as ‘Coloured’, make up 8.8% of the SA population (Statistics South Africa, 2019). SAMA are a phenotypically heterogenous group with highly variable, mixed socio-cultural ancestry, arising from genetic mixing between indigenous inhabitants of the Cape (Khoesān) and European, African, and Asian imported slaves and migrants (Adhikari, 2006; De Wit et al., 2010; Isaacs-Martin & Petrus, 2012; Inwood & Masakure, 2013; Montinaro et al., 2017).

Many metric sex estimation methods demand ancestry as a prerequisite and, these methods are often specific to only SAAA or SAEA (Asala, 2001; Asala, Bidmos & Dayal, 2004; Dayal &

Bidmos, 2005; MacAluso, 2011). More recent attempts have been made to study SAMA (Mokoena et al., 2019; Tawha et al., 2020) but when ancestry is unknown, options become limited.

2.3 AGE-AT-DEATH

Bone modification exists as a multidimensional amalgamation of skeletal maturation maintenance and degradation over the course of any individual's lifetime. Processes associated with aging affect all elements of bone from the periosteal surfaces of long bones to bone mineralisation of cortical bone and the medullary cavity (Vance, 2007).

Net gain in bone mass continues after puberty in women until the late 20s (28.0-29.5) with dietary intake of calcium, physical exercise and the consumption of oral contraceptives playing a role in enhanced net gains (Recker et al., 1992).

Mechanical loading and strain are highly influential in the rate of bone deposition, muscle wasting and bone loss (Burr, 1997). Remodelling occurs constantly in response to stressors and as part of the aging process. Net loss at the periosteal surface relative to new bone deposition at the cortical-endosteal surface begins before midlife and accelerates with advancing age (Compston et al., 2007). Net cortical bone loss in individuals of advancing age at the cortical-endosteal surface causes expansion of the medullary cavity and a reduction in cortical wall width (Seeman, 2003). Male and female bone resorption rates are somewhat equal; however, males form new bone (periosteal deposition) at a greater rate than females. Net bone loss in females therefore exceeds males and is likely a ramification of hormonal changes associated with the menopausal transition (Hernandez, Beaupré & Carter, 2003).

In response to cortical bone thinning, midshaft diameter has been shown to rise with increasing age (Smith & Walker, 1964). Biomechanical remodelling of bone leading to increased shaft diameter results in improved resistance to bending and torsion in limb shafts and at joint surfaces thus providing mechanical resistance to reduce fracture risk (Szulc, 2006). This increase in shaft diameter is more pronounced in men due to their continued higher rate of periosteal bone depositions as they age compared to females (Seeman, 2003).

An increase in skeletal measurement values with advancing age-at-death is therefore expected in both males and females due to bone remodelling and adaptation to cortical bone loss.

2.4 BILATERAL ASYMMETRY

Along with sex, age-at-death and ancestry, bilateral asymmetry is a factor which may contribute to skeletal variation. The most influential factor in bilateral asymmetry, the size- and shape -based differences between the left- and right- sided bones of the skeleton, is differential biomechanical stress during bone growth (Dillon, 2014). When a mechanical force acts on a bone, it remodels in response to the stress (Lazenby, 2002). Therefore, long term, repetitive mechanical loading and contralateral muscle contractions may lead to skeletal asymmetry, specifically cross-symmetry in bone length, muscle bulk, cortical dimensions, bone weight and articular dimensions (Kanchan et al., 2008). Kanchan et al. (2008) state that the upper dominant limb is larger and more robust than the non-dominant one and the contralateral lower limb to the dominant hand is more robust.

The upper limb manifests more skeletal asymmetry than the lower limb (Latimer & Lowrance, 1965). Loading and disuse in the upper limb and bipedal locomotion in the lower limb are credited for causing bilateral asymmetry to be more pronounced in the upper limb (Lazenby, 2002). The numbers vary, but approximately 90% of all people are thought to be right-handed (Jung & Jung, 2009). Thus, the dominant, and therefore more robust, upper limb should tend to be the right one.

Plochocki (2004) observed bilateral asymmetry of limb articular surfaces. They found that right limb measurements were larger than left in approximately 50% of cases for the upper limb but lower limb measurements were more often equal for left- and right-sided bones (Plochocki, 2004). These findings are consistent with the Kanchan et al. (2008) expectation that the dominant upper limb would be larger, and that bilateral variation would be less obvious in the lower limb. Carvallo & Retamal (2020), in contrast to Plochocki (2004), concluded that there were no significant differences between the left- and right-sided bones in their study sample.

2.5 RELATIVE PRESERVATION OF SKELETAL ELEMENTS

In attempting to solve the sex classification problem using skeletal remains, there are a litany of plausible skeletal elements and combinations which can be selected. Important considerations include: the relative rate of survival (preservation) of a skeletal element and, the level of sexual dimorphism presented by that element (Bidmos, Gibbon & Štrkalj, 2010; MacAluso, 2011; Šlaus et al., 2013).

Recovered human remains may show signs of resultant damage from various taphonomic processes. Factors influencing preservation include: environment (Megyesi, Nawrocki & Haskell, 2005; Donaldson & Lamont, 2014), burial (Dent, Forbes & Stuart, 2004), mass graves (Haglund, Connor & Scott, 2001), vertebrate and invertebrate scavenging (Spies, Gibbon & Finaughty, 2018), scavenger-induced scattering (Spies, Finaughty & Gibbon, 2018), dismemberment (Konopka et al., 2007), cannibalism (Degusta, 1999), and entomology (Benecke, 2015).

Consequently, forensic anthropologists may be restricted from examining certain features or elements more commonly used for sex estimation. There is thus merit in developing methods based on many bones, and particularly those shown to have the best expected survival and recovery rates.

Bone preservation patterns have shown to be fairly consistent across different recovery sites regardless of varying taphonomic processes (Stojanowski, Seidemann & Doran, 2002). Less dense, more porous bones such as the ribs, sternum and vertebrae show poor preservation whilst denser areas including the long bones have proven to survive better (Willey, Galloway & Snyder, 1997). The cranium is the most widely recovered skeletal element, however, the facial bones are commonly damaged or missing (Spennemann & Franke, 1995; Stojanowski, Seidemann & Doran, 2002).

Traditionally, the skull and pelvis are most used to infer sex (Buikstra & Uberlaker, 1994). However, due to various taphonomic and anthropic processes, the preservation of these elements may render them unusable (Asala, Bidmos & Dayal, 2004; Simmons, Jantz & Bass, 2015). In these cases, other skeletal elements must be utilized.

The long bones display high levels of sexual dimorphism (Spradley & Jantz, 2011) and present better preservation than both the skull and pelvis due to their robusticity and tubular shape (Soni, Dhall & Chhabra, 2010). Complex joints, such as the shoulder, hip, elbow, and knee have also been shown to exhibit sexual dimorphism (Asala, Bidmos & Dayal, 2004; MacAluso, 2011; Vance, Steyn & L'Abbé, 2011; Vance & Steyn, 2013). Whilst the scapula is generally not well preserved due to its thin, irregular shape; the glenoid fossa, which represents a complex joint, is preserved in 62% of cases (Stojanowski, Seidemann & Doran, 2002).

In light of the information presented in this section, the literature reviewed in the following sections will focus on the glenoid fossa, head of the humerus, distal humerus, acetabulum, head of the femur, distal femur, and proximal tibia.

2.6 SEX ESTIMATION METHODS

As introduced in Chapter 1, sex estimation can be performed using a host of methods. In recent years molecular methods, utilising DNA analysis, have gained popularity but historically, sex estimation techniques including morphological, metric and more recently, geometric morphometric methods were favoured. Each of these approaches has its own assortment of efficacies and failings which will be explored.

2.6.1 MOLECULAR TECHNIQUES

Molecular sex estimation methods use DNA extracted from hard tissue (bones or teeth) to determine sex. A specific gene or genes in the sex chromosomes is isolated and amplified before being scrutinised to find pertinent genetic information which can be used to assess sex (Latham & Miller, 2019). The genetic elements most used in sex estimation are zinc finger proteins and amelogenin genes.

DNA extracted from teeth was used by (Pillay & Kramer, 1997) to evaluate sex by amplifying the zinc finger protein gene using PCR. Unfortunately, PCR is inhibited by heavy metals and is temperature sensitive (Urbani, Lastrucci & Kramer, 1999). This is problematic given that human remains in forensic and archaeological cases are often exposed to soils which contain heavy metals.

Novel methods have been developed in South Africa which, instead of PCR, use silica extraction followed by isolation of the amelogenin gene (Gibbon, Paximadis, et al., 2009). These new silica extraction amelogenin isolation methods were tested on human remains and demonstrated to work well as they relieve the drawbacks associated with PCR (Gibbon, Paximadis, et al., 2009). Amelogenin is also useful for sex assessment in fragmentary, fire damaged, and immature remains where traditional sex estimation methods cannot be applied (Faerman et al., 2000; Gibbon, Paximadis, et al., 2009). Urbani et al. (1999) assessed molecular degradation, due to high temperatures, in human teeth and found that DNA from fire victims remains intact and viable for sex identification.

A significant shortcoming of molecular methods is that they are often destructive, requiring hard-tissue samples to be milled into a fine powder for DNA extraction. Skeletal collections typically aim to preserve their sample material, thus making it difficult to access skeletal material for DNA analysis. For this reason, (Woodward, Penny & Ruff, 2006) and (Woodward, Penny & Ruff, 2006; Gibbon, Penny, et al., 2009) sought to develop a minimally invasive sampling procedure by using the intercondylar fossa of the femur. This technique preserves the physical properties of the skeletal remains therefore retaining their usefulness for metric, morphological or morphometric analyses and eliminating the shortcomings generally associated with sampling. Unfortunately, this method is only useful when the femur is available for sampling.

The many merits of molecular sex estimation include: their high level of accuracy; ability to determine sex regardless of fragmentation or missing skeletal elements; and, importantly, their ability to determine sex even when the remains belong to a minor or individual of unknown ancestry, nationality, or chronologic age (Krishan et al., 2016). Despite these advantages, the nature of molecular analytics makes these techniques expensive, specialized, and time-consuming. For these reasons, molecular approaches are currently not easily accessible or user-friendly and cannot be applied quickly in the field for archaeological or forensic cases.

An alternate sex estimation method (which is speedy, non-destructive, and economical) is morphological sex estimation.

2.6.2 MORPHOLOGICAL OR NON-METRIC

The visual observation of sexually dimorphic bony landmarks is a popular sex estimation technique due to quick and easy application for preliminary assessment and requiring no additional specialised equipment. Visual observation is however highly subjective and reliant on observer experience (Walrath, Turner & Bruzek, 2004).

In the late 1990s, most morphological standards for human identification used in SA were adapted from international sources and not developed specifically for the local population (Steyn, Meiring & Nienaber, 1997). In response, several studies over the last 20 years have focussed on testing existing methods and developing new morphological standards for South Africans.

The skeletal elements most commonly used in morphological sex estimation are the skull and pelvis although other bones, including the humerus and femur, may also be used, these have not been widely studied in SA (Christensen & Crowder, 2009).

The cranium has been extensively studied in South Africa (De Villiers, 1968; Loth & Henneberg, 1996; Kemkes-Grottenthaler, Löbig & Stock, 2002; Balci, Yavuz & Cağdir, 2005; Oettlé, Pretorius & Steyn, 2005; Seedat, van Niekerk, et al., 2009a) however, this is not relevant to the current study which will not examine the skull.

The human pelvis, often considered to be the most reliable skeletal element for sex assessment, was adapted to accommodate parturition in females which made it highly sexually dimorphic (Lundy, 1998; Steyn, Pretorius & Hutten, 2004). Female pelvic morphology is a product of evolutionary selection, adaptation, and developmental plasticity in reaction to ecological and nutritional factors (Huseynov et al., 2016). When assessing a collection of morphological traits on 400 known SAAA and SAEA skeletons, it was found that pubic bone morphology, as introduced by (Phenice, 1969) and later modified (Lovell, 1989; Sutherland & Suchey, 1991), best feature in both sex and ancestry groups with 88% accuracy of sex estimation in SAEA and 84.5% in SAAA (Patriquin, Loth & Steyn, 2003). Interestingly, Patriquin et al. (2003) note that the ischiopubic ramis and greater sciatic notch were unexpectedly inferior in accurately estimating sex.

So, morphological methods are a useful tool for quick preliminary assessment of sex in the field, but they are exceedingly subjective and worryingly reliant on the observer's experience and expertise (Lovell, 1989; Steyn, Pretorius & Hutten, 2004).

A less subjective alternative to morphological methods may be geometric morphometrics.

2.6.3 GEOMETRIC MORPHOMETRICS

Geometric morphometrics, the quantification of morphological traits in terms of both size and shape, is valuable in reducing the subjectivity associated with morphological sex estimation and has proven to be highly accurate on traditional skeletal elements, like the pelvis, with 100% accuracy when using 36 3D landmarks (Bytheway & Ross, 2010).

In response to Patriquin et al. (2003), who, as mentioned earlier, found the greater sciatic notch to be an inferior classifier of sex, Steyn et al. (2004) used geometric morphometrics to assess the greater sciatic notch in 115 South African skeletons. They found that whilst females tend to have characteristically wide notches (Buikstra & Uberlaker, 1994), SAEA males tended to be highly variable, with shapes scattered across the full range (Pretorius, Steyn & Scholtz, 2006). Similar findings were reported with a non-SA population by (Walker, 2005), who also noted that age may be a significant factor owing to classification accuracy when using the pelvis. The greater sciatic notch was further assessed in SAAA using the TSP series of programs for geometric morphometric analysis and reported sex prediction accuracies of 93.1% for males (Pretorius, Steyn & Scholtz, 2006), although the sample size was notably small (31 males, 29 females). This confirms that that sciatic notch can be useful for males of SAAA ancestry, despite being unreliable for SAEA males.

The scapula has been assessed using 2 very different geometric morphometric techniques in South Africa. Scholtz, Steyn & Pretorius (2010) used photographs of 90 SAAA (45 male, 45 female) to estimate sex by manually assigning 21 landmarks and passing the data through analytic software. Although time consuming, and requiring experience in assigning landmarks, they were able to assign sex with 91-95% accuracy (Scholtz, Steyn & Pretorius, 2010). Alternatively, Macaluso (2011) studied the glenoid fossa of 120 SAAA (60 male, 60 female) from the Pretoria Bone Collection. Image analysis of photographs was used to calculate height, breadth, area, and perimeter of the glenoid cavity. The best predictor of sex was glenoid area with 88.3% followed by glenoid breadth (which can also be metrically assessed using sliding callipers), yielding an 85.8% prediction accuracy. It was further noted that multivariate models were unable to provide increased accuracy over univariate models (MacAluso, 2011).

A study of 330 SA humeri was able to classify sex with accuracies between 78% and 91% (Vance & Steyn, 2013). This confirmed that geometric morphometrics can quantify 3D landmarks such as olecranon fossa shape, angle of the medial condyle and trochlear symmetry to estimate sex with the same or better accuracy than visual observation alone, which achieved accuracies between 74% and 90% (Vance, Steyn & L'Abbé, 2011). The humerus was further studied in a large SA sample of 1046 individuals which claimed that including ancestry during sex estimation had the power to improve classification accuracy (Maass & Friedling, 2019). The improvements were, however, minor with the overall accuracy of sex estimation being 80.7% without ancestry, compared to 82.4% in SAAA, 79.6% in SAMA and 81.1% in SAEA when ancestry was accounted for (Maass & Friedling, 2019).

Whilst geometric morphometrics provide a means for accurate sex classification, problems arise when using these techniques in the field as very specific positioning of bones, as well as exact camera placement, is usually required, and software is often needed to perform the time-consuming process of assigning landmarks and co-ordinates (Christensen & Crowder, 2009). However, with future improvements in technology and accessibility, it has the potential to eradicate subjectivity whilst maximising the benefits of metric and non-metric sex estimation techniques.

To fully appreciate the impact that geometric morphometrics might have in the future, we must first properly explore the benefits of metric methods.

2.6.4 METRIC METHODS

Some parts of the human skeleton fail to present reliably sexually dimorphic morphology. In these instances, when sexual differences are largely size-based, and when incomplete or fragmentary remains are found, the observer may be forced to utilise less sexually dimorphic skeletal elements, to which metric methods may be applied. Metric methods involve single or multiple mensurations which are substituted into mathematical functions or models to estimate sex. Metric sex estimation methods are customarily advantageously easy to apply in the field, their accuracy is quantifiable, their results are usually simple to interpret regardless of expertise, and observer error as well as replicability are easily assessed.

For the purposes of sex classification, metric mensurations are generally subjected to discriminant function analysis (DFA). First introduced by Fisher in 1936 (Fisher, 1936), the general premise of DFA is to classify an observation into 1 of 2 or more well-defined groups. Differences between group centroids are determined and studied to determine the contribution of multiple variables to inter-group separation and the costs of making an error as well as the probabilities of an observation belonging to one of the groups (Büyüköztürk & Çokluk-Bökeoğlu, 2008). Allocation rules, optimised based on these criteria, are then set up to assist in assigning an individual to a particular group or response variable. These rules are combined to create a discriminant function. The resulting discriminant function is then used as a linear classifier which attempts to express the response variable (in this case sex) as a linear combination of multiple predictors. DFA will be further explained in a subsequent section.

As previously mentioned, the femur and pelvis are often well-preserved in the medicolegal setting due to their generally robust structure and relatively larger size. As such, these bones, as well as their associated joint cavities (acetabulum and glenoid), have been studied to determine their value for metric sex estimation.

Patriquin et al. (2005) evaluated the pelvis of SAAA and SAEA individuals, selecting 6 sexually dimorphic mensurations for use in discriminant analysis. A multivariate model using all 6 mensurations was able to classify sex with 90-98% accuracy, whilst acetabulum diameter alone was on average only 82% accurate (Patriquin, Steyn & Loth, 2005). In cases of highly fragmented remains, where only the acetabulum remains intact, a second measurement of acetabulum diameter has been shown to be useful. Transverse acetabulum diameter, as presented in Bubalo et al. (2019), in conjunction with acetabulum diameter classifies sex consistently with 88% accuracy (Bubalo et al., 2019). However, when this discriminant function is adjusted for and applied to a SA sample, accuracy declines to 82.1% (Scott, 2019). This illustrates the geographic specificity of metrics and highlights the need for validation of international methods on the local population.

Asala (2001) estimated sectioning and demarcation points for 2 mensurations of the femoral head, which were then later validated to verify their reported accuracies (Asala, 2001, 2002). Asala (2002) defined a male and female sectioning point and reported that whilst all individuals above the male and below the female sectioning points were accurately classified, those between the male and female sectioning points could not be identified, leaving 68% of all individuals in the study unidentified. These findings illuminate the large degree of overlap between male and female measurement values and confirm that the use of simple demarcation and sectioning points is not sufficient to secure an accurate classification of sex.

The usefulness of the fragmentary femur for sexing was studied by Asala et al. (2004) using 5 proximal and 3 distal femoral mensurations. These mensurations were subjected to discriminant function analysis. In contrast to the low classification accuracy (32%) achieved using sectioning and demarcation points, univariate analysis attained a 68-83% correct classification rate whilst multivariate analysis proved to be superior, attaining accuracies of 83-85% (Asala, Bidmos & Dayal, 2004). Asala et al. (2004) conclude that they believe either the femoral head or the distal joint surface of the femur can be used in isolation to correctly estimate sex. This study highlights the necessity for statistical methods with the ability to separate the sexes.

The tibia has only been studied in conjunction with the femur within South Africa. The tibia was first examined in SA by Steyn and Işcan (1997) using a relatively small sample of 106 SAEA. The distal portion of the tibia and femur were reported as being most accurate for sex classification with discriminant functions being 86-91% accurate when using either one bone, or both (Steyn & Işcan, 1997). Alone, tibia proximal breadth classified 86.8% of individuals correctly. This is higher classification accuracy than Asala et al. (2004) reported for any femoral univariate model.

As previously discussed, the glenoid cavity of the scapula was examined by MacAluso (2011). They used geometric morphometric techniques to collect and analyse data. However, the mensurations which they measured using software can be manually measured using sliding callipers and are thus included in this section. Discriminant functions generated for glenoid height and breadth were able to estimate sex with ~88% accuracy for SAAA (MacAluso, 2011). A concern with these results is that they were neither validated on an independent sample nor using any form of cross-validation. Validation of sex prediction models is particularly important as the reported prediction accuracies, which are often determined from the same sample used to derive the model, may represent the outcomes of an overfitted model (Roelofs et al., 2019).

The humerus has proven to be a reliable skeletal element for sex estimation with similar findings from 2 authors. Steyn & Işcan (1999) studied 101 SAEA and 88 SAAA individuals and developed stepwise discriminant functions for sex estimation as well as demarking points. The stepwise multivariate model estimated sex with the same accuracy as epicondylar breadth alone in females; whilst the accuracy of the univariate model was significantly lower in males. Robinson & Bidmos (2009) tested the functions developed by Steyn & Işcan (1999) on a cohort of 264 SAEA humeri from UCT, Pretoria and Dart bone collections. Whilst accuracies varied between collections with the largest differences being between UCT and the 2 Gauteng-based collections (probably due to regional differences), the overall average accuracy of ~92% was in agreement with the original reported accuracy of 92.5% (Steyn & Işcan, 1999; Robinson & Bidmos, 2009).

All of these papers show the usefulness of metric methods for sex estimation but, as with all sex estimation methods, there are several downfalls to using metric methods. These include population specificity of discriminant functions (Bidmos & Dayal, 2004; Steyn & Patriquin, 2009; Dillon, 2014; Curate et al., 2016; Kotěrová et al., 2017) and difficulty in overcoming the overlap in measurement values between males and females (Asala, 2002).

In addition to the sectioning points, demarcation points and discriminant analysis traditionally employed by scientists, machine learning (ML) algorithms have more recently been used in solving sex classification problems. They are, for all intents and purposes, the superset of algorithms of which traditional metric methods are a subset of, and so are a natural evolution of the research.

2.7 MACHINE LEARNING IN BIOLOGICAL ANTHROPOLOGY

Within the sphere of metric sex estimation, ML has emerged as a new player within the last 10 years. ML methods usually do not require training data to meet a particular set of statistical assumptions, such as normality, as in some traditional methods. The main benefits of ML algorithms are that they can more exhaustively assess the training data and iteratively solve optimization problems to find, often non-linear, solutions with the best possible classification rates; and that ML methods are not merely reliant on group parameters such as mean and covariance. Another benefit of ML models is that, in most cases, they require no significant expertise to use once trained, and are rapid to employ for sex prediction (Seedat, van Niekerk, et al., 2009b).

Outside of SA, ML has been applied to geometric morphometrics using computed tomography (CT) scans and photographs. Bewes et al. (2019) applied GoogLeNet neural networks to develop sex prediction methods for crania. They used a sample of 900 CT scans from Australia's Royal Adelaide Hospital PACS database and were able to predict sex with an accuracy of 95%. Toneva et al. (2020) also studied the cranium, sampling 393 individuals, and using ML techniques to develop classifiers. This study assigned landmarks rather than allowing software to find patterns in the images and were able to achieve a maximum prediction accuracy of 91.9% using a less complex reduced error pruning approach, JRIP (Toneva et al., 2020). These 2 papers show that although a more complex model may be able to predict sex with higher accuracy, much simpler models still have the power provide highly accurate predictions.

More relevant to this study is the femur. The femur has proven to be a useful tool for sex classification within the field of machine learning with ML methods having been applied to geometric morphometric data (Curate, Albuquerque, et al., 2017) and metric data (Curate et al., 2016; Curate, Umbelino, et al., 2017; Coelho & Curate, 2019; Carvallo & Retamal, 2020). A series of geometric morphometric sex classifiers were trained utilising the femur. Using a Portuguese sample of 224 densitometric femoral scans, Curate, Umbelino et al. (2017) reported accuracies ranging from 90.0-92.9% using a variety of ML algorithms. However, linear discriminant analysis (LDA), logistic regression (LR) and REPTrees (Reduced Error Pruning), shared a prediction accuracy of exactly 91.7% when applied to an independent validation sample (Curate, Albuquerque, et al., 2017). This, and the small (<3%) difference in accuracy between all applied algorithms, emphasizes the need for a large enough holdout sample to avoid convergence of error rates and limit the effects of outliers within the holdout.

Using a metric approach, ML techniques were applied to pelvic mensurations from 256 Portuguese individuals. The highest prediction accuracy achieved was 97% using a colossal 38 variables (Coelho & Curate, 2019). This study made use of a number of different machine learning algorithms including: partial least squares regression analysis, which is a supervised dimension-reduction technique originally developed for used in regression problems; simple neural networks, adaptive models that draw inspiration from non-linear learning occurring in neuronal networks of animal brains; decision trees, which follow a series of tiered tests with IF-THEN logical rules; random forests, ensemble methods making use of multiple decision trees; boosted logistic regression, a set of ensemble algorithms which sequentially apply additive models before introducing a cost function; K nearest-neighbour, an instance-based classifier that stores the training set and then compares new, uncategorized records by comparing them to similar records within the training set; Naïve Bayes, which consists of a set of probabilistic classifiers which use Bayesian statistical rules (i.e. conditional-independence assumptions) to predict the probability of an event; LDA which relies on the combination of multiple linear predictor variables which split mutually exclusive groups (in this case male and female) and derive discriminant functions which typify the differences between groups; and fuzzy trees, which are used to describe the membership degree of an instance following a Gaussian curve (Coelho & Curate, 2019).

Coelho & Curate (2019) noted that whilst using all 38 variables could produce accuracies of 83-97%, despite having the best results, similar performance metrics can be achieved by measuring only 3 variables. They raise the question as to whether the slight improvement in accuracy is worth the time taken to complete the extra 35 measurements. If measuring 3 variables will give you 92% accuracy, would you measure 38 to achieve 96% accuracy?

Continuing with the metric approach, the femur has been comprehensively studied in a European population and a South American population. The most reliable algorithm in the Chilean study was logistic regression with a reported accuracy of 92.9% for a univariate model and up to 95.7% for multivariate (Carvallo & Retamal, 2020). In comparison, also using logistic regression, the Portuguese study reported sex prediction accuracies not exceeding 87.5% for univariate models and not exceeding 92.5% for multivariate models (Curate, Umbelino, et al., 2017). Perhaps these differences were due to population differences (like homogeneity). Curate, Umbelino, et al. (2017) noted that they found the ML classification method chosen did not greatly affect the accuracy or bias associated with the model, so it is unlikely that the discrepancies are due to different statistical approaches.

A machine learning study exists which made use of data collected from South African skeletal collections.

Mircea (2016), a Romanian paper, assessed 19 metric traits from the ulna and radii of 400 SAAA individuals which were sampled for a University of Pretoria Masters' thesis (Barrier, 2007). This paper focused on using Fuzzy decision trees in the hopes of overcoming populational dependencies associated with metric sex estimation by factoring in "populational particularities" to assist in correctly identifying individuals which fall into the area of overlap between males and females. The reported prediction accuracies were 84-89%, with Fuzzy trees performing better than traditional linear discriminant functions (Barrier, 2007; Mircea, 2016). These findings show that ML methods can outperform the most commonly used modelling technique in South Africa.

This selection of studies shows that ML models have the capacity to classify sex with high levels of accuracy, certainly competitive with all other sex estimation methods. Very little work has however been performed on a South African sample using methods aside from discriminant analysis. This thesis hopes to rectify this gap in the current literature and demystify machine learning as a sex classification tool.

CHAPTER 3: MATERIALS

3.1 SKELETAL COLLECTIONS

In South Africa, most documented human skeletal remains in skeletal repositories are accessioned from donated cadavers. The Inspector of Anatomy is responsible for ensuring that cadavers are obtained in an ethical and legal manner and is mandated to carry out regular inspections of anatomy departments and their specimens (Pillay, McQuoid Mason & Satyapal, 2017). Cadaveric remains, used for teaching purposes, are those of individuals whose bodies have been bequeathed to scientific research or have remained unclaimed by their kin after death (Gangata et al., 2010). Following dissection by students, some cadavers are macerated each year and added to university skeletal repositories. In SA, cadaveric skeletal material may be accessed, with permission, for scientific study from the UCT Cadaveric Skeletal Repository (Maass & Friedling, 2019), the Stellenbosch University (SUN) Kirsten Collection (Alblas, Greyling & Geldenhuys, 2018) the WITS Raymond A. Dart Collection (Dayal et al., 2009) and the University of Pretoria (UP) Bone Collection (L'Abbe, Loots & Meiring, 2005).

The UCT Cadaveric Skeletal Repository and WITS Raymond A. Dart Collection are utilized in this study. The composition of cadaveric skeletal remains in these collections is summarised in Table 3-1

3.1.1 UNIVERSITY OF CAPE TOWN CADAVERIC SKELETAL REPOSITORY

The UCT Human Skeletal Repository is housed by the Division of Clinical Anatomy and Biological Anthropology within the Health Sciences Faculty at the University of Cape Town. This collection includes archaeological, cadaveric, and forensic remains. The UCT Repository is relatively small in comparison to other skeletal collections in SA. It is comprised of approximately 350 individuals of known age, sex, and ancestry. The majority (206/346) of the repository are bequeathed individuals of European descent with an age-at-death exceeding 60 years. The balance of the collection is composed of younger, 'unclaimed' individuals (largely SAAA or SAMA) from either mortuaries or state hospitals in the Cape Town Metropole and surrounding areas within the Western Cape (da Silva, 2006; Maass & Friedling, 2019)

3.1.2 UNIVERSITY OF THE WITWATERSRAND RAYMOND A. DART COLLECTION

The Raymond A. Dart Collection, colloquially called the 'Dart Collection', is the oldest and largest skeletal collection in SA. The Dart Collection can be found at the School of Anatomical Sciences at the University of the Witwatersrand. This collection comprises approximately 2600 individuals of which the majority are SAAA and includes several non-South Africans which were excluded from the current study. The collection consists of proportionally more males (71%) than females and has a mean age-at-death lower than 60 years. All cadavers collected before 1958, as well as a large proportion of individuals in later years, were derived from unclaimed bodies in regional state hospitals within the Gauteng province. As such, some demographic details are estimates. In 1959, disaster struck the collection in the form of a flood which led to some of the skeletons being mixed up. It is thus advisable to avoid pre-1959 skeletons (Dayal et al., 2009).

Table 3-1 Composition of skeletal remains in the UCT and Dart skeletal repositories divided by individuals with an age-at-death below and over the age of 60 years.

		UCT		**Dart**	
	Sex	<60	>60	<60	>60
EASA	Male	14	116	68	192
	Female	10	90	40	168
SAAA	Male	12	14	977	333
	Female	5	1	362	101
SAMA	Male	24	26	42	29
	Female	24	10	32	8

3.2 ETHICAL AND DEMOGRAPHIC CONSIDERATIONS

When using a cadaveric skeletal sample, there are several limitations and inherent biases which must be considered before drawing conclusions or making comparisons with the living population. Cadaveric human skeletal remains have been added to SA skeletal repositories for over a century meaning that the year-of-birth for these individuals may be as early as the late 1800s. The demography of the collections in terms of age-at-death, sex, ancestry, year-of-death, and socioeconomic status are not perfectly analogous to the living population and have been influenced by institutional acquisition practices over the last 150 years.

Skeletal collections/repositories are commonly used as study samples in forensic anthropology; however, demographic biases have been introduced into these collections as a result of public perception of body donation, along with religious and cultural viewpoints around death, burial, and the afterlife (Grivas & Komar, 2008). A proportion of cadavers in SA are derived from cadaver donation programs (Gangata et al., 2010; Gangata, 2015) whilst the balance of human remains is donated to universities by the state. State donated remains are a bone of contention. Arguments for and against the study of state donated remains exist, with a general movement away from the use of skeletal remains which have not been personally bequeathed (Gangata et al., 2010; Gangata, 2015; Cornwall, Callahan & Wee, 2016; Jayakumar, Athar & Ashwood, 2020). In an interview, an Inspector of Anatomy stated that prior to the passing of the National Health act in 2003, unclaimed and unidentified bodies were the property of the state and instead of a pauper burial, these bodies were often donated by Inspectors of Anatomy to departments of anatomy (Pillay, McQuoid Mason & Satyapal, 2017). However, this practice goes against the religious and cultural beliefs of many South Africans.

African communities have expressed objections towards body, tissue, and organ donation, citing concerns about 'ancestor reverence' (Zulu, 2013; Makgahlela et al., 2021) and the belief that the body and soul must remain together after death (L'Abbe, Loots & Meiring, 2005). Similarly, some Muslim communities are generally not inclined to participate in body or organ donation due to the potential influence on one's afterlife and advice from religious leaders despite unclear directives in religious texts (Rokade & Gaikawad, 2012; Gürses et al., 2019; Ali et al., 2020). Therefore, these groups are more likely to be underrepresented in skeletal repositories.

It is often assumed that most SAEA cadavers are bequests from elderly middle to upper class individuals, however between 1963 and 2005, 50.07% of all SAEA at UCT were bequests whilst the remaining 49.93% were state donated (da Silva, 2006). Over the same period (1963-2005), only 2.3% of all bequests at UCT were from SAMA (da Silva, 2006). This corroborates the assumption that the majority of SAMA and SAAA remains were state donated. In actuality, between 90 and 100% of all non-SAEA individuals in university skeletal repositories were state donated (Dayal et al., 2009; Alblas, Greyling & Geldenhuys, 2018; Maass & Friedling, 2019). This pattern of incongruent donation is also prevalent at Stellenbosch University, where between 1956 and 1996, 99% of SAAA, 95.8% of SAMA and only 34.1% of SAEA cadavers were state donated (Labuschagne & Mathey, 2000). These statistics are not current or up to date, having been published 15 and 25 years ago, but individuals from these periods remain in skeletal repositories today.

3.3 ETHICAL CLEARANCE

The National Health Act No. 61 of 2003 governs the use of cadaveric skeletal remains for research purposes in SA. The study of human skeletal remains at an academic institution generally falls under this act and is monitored by the Inspector of Anatomy. A submission was made to the UCT Human Research Ethics Committee to ensure that the research was deemed ethical. The UCT HREC REF number for this study is 838/2020sa. Applications were also made to each of the respective skeletal collections for permission to collect data.

3.4 SAMPLE DEMOGRAPHIC

The final sample included 325 males and 325 females for a total of 650 South African individuals. This represents the South African population of 58.78 million (Statistics South Africa, 2019) with a confidence level of 95% and a margin of error of 3.84%. This sample was selected to account for incomplete skeletons and include a validation sample. The ancestry of individuals in this sample is approximately demographically representative in line with the 2019 SA population estimates (Statistics South Africa, 2019). Table 3-2 shows the composition of skeletal remains in the study sample.

Table 3-2 Summary of study sample.

	Sex		
Repository	Male	Female	**Total**
UCT	84	40	**124**
Dart	241	285	**526**
Total	**325**	**325**	**650**

3.4.1 EXCLUSION CRITERIA

Exclusion criteria including age, ancestry, trauma, and pathology are specified below.

3.4.1.1 AGE

Individuals born before 1900 were excluded from the sample to reduce the effects of secular trend and create the most modern and population representative sample possible within the constraints of skeletal availability. Individuals below the age of 18 and above the age of 65 were also excluded due to the impact of aging on bones (Vance, 2007) and relevance for forensic application (Evert & Rossouw, 2011; Herbst, Tiemensma & Wadee, 2015).

3.4.1.2 ANCESTRY

Only individuals of SAAA, SAEA and SAMA were included in the study cohort. Foreign individuals and those from other minority groups were excluded due to underrepresentation in skeletal collections. For example, an Indian cohort was excluded as only 5 individuals are present in the Dart collection and none are reported in the UCT repository (Dayal et al., 2009; Maass & Friedling, 2019).

3.4.1.3 TRAUMA

All skeletal elements presenting signs of antemortem or perimortem trauma, including fractures and corrective surgery, were excluded. Additionally, bones with post-mortem damage and deterioration were assessed and excluded if likely to hinder accurate measurement.

3.4.1.4 PATHOLOGY

All elements with markers of pathology were not sampled. Additionally, adjacent bones (i.e., pelvis – femur) were also excluded if there was joint pathology or trauma.

3.4.1.5 INCOMPLETE SKELETONS

Incomplete skeletons were not excluded. Data from incomplete skeletons can still be useful for univariate analyses.

CHAPTER 4: METHODS

4.1 DATA COLLECTION

Individuals were preselected before measurements were taken with the intention to build a sample with well-distributed age-at-death, most recent possible year-of-birth, equal sex distribution and a demographically representative proportion of individuals from each ancestry group. The ancestry of individuals in this study was not recorded.

Demographic details, specifically: skeleton number, age-at-death, year-of-death, and sex for the selected individuals were imported directly into a custom interface stored in a PostgreSQL database (PostgreSQL 2.4, PostgreSQL Global Development Group) before sampling. The GitHub repository (containing the source code) for this database can be found at: https://github.com/mscott1037/the_bone_collector.

This allowed sampling to be performed blind as individuals could be retrieved from the database by entering only the skeleton number during data collection.

The mensurations described in Table 4-1 were measured using ACCUD 150mm digital sliding callipers (see Appendix A calibration certificate). Each measurement was performed 3 times and entered directly into the custom database. Bone diagrams in Figure 4-1 visually define each of the parameters. This same set of mensurations were investigated in the author's Honours thesis (Scott, 2019).

To assess inter- and intra-observer error: the initial observer, as well as an independent observer (a PhD student from UCT) resampled a randomly selected group of individuals.

After data collection, data were sanitised and imported into MATLAB for statistical analysis.

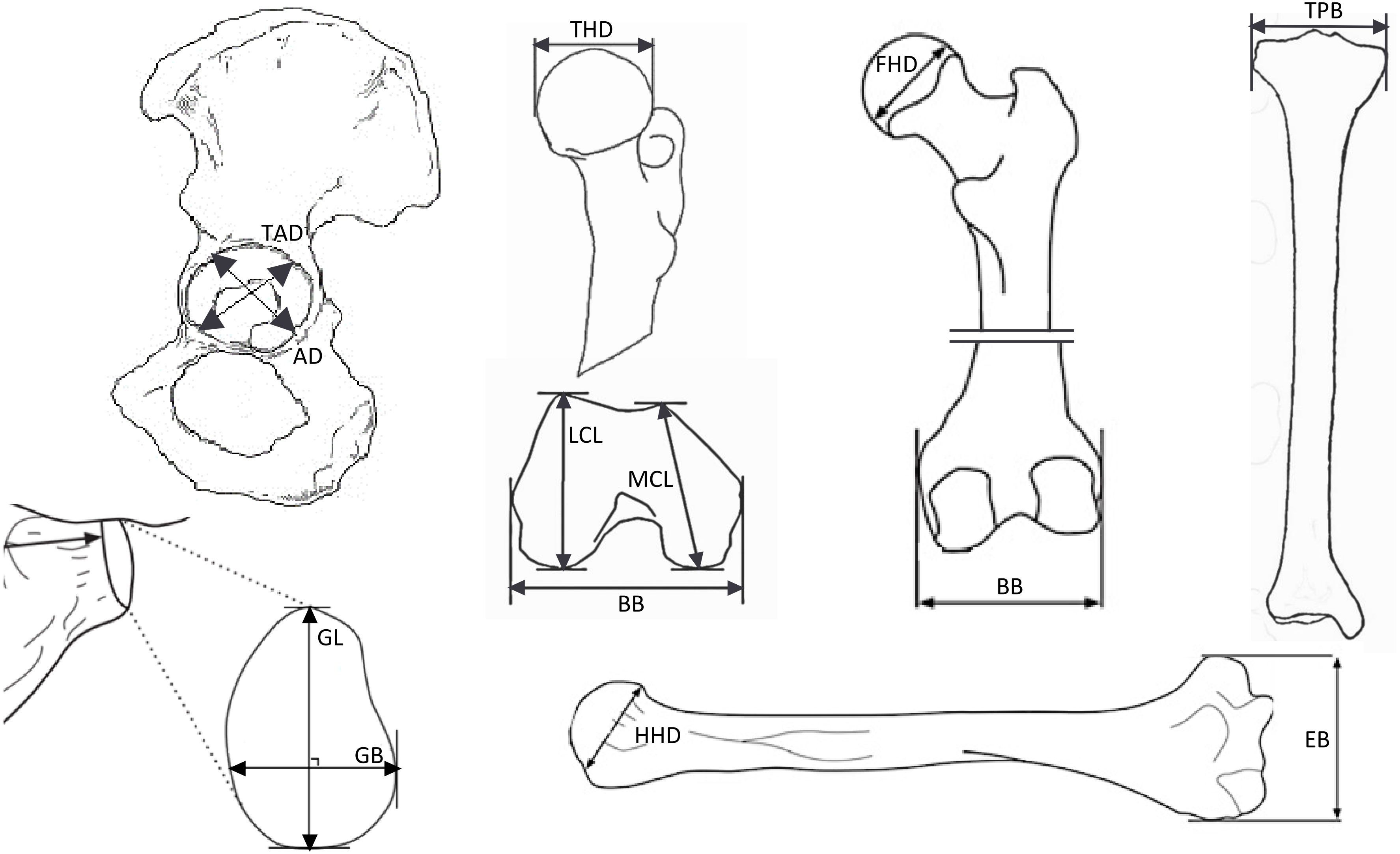


Figure 4-1 Diagrams showing mensurations for each bone.

Table 4-1 Descriptions of parameters to be measured on the proximal and distal elements of the long bones and their associated joint cavities.

	Parameter	**Description**	
AD	**A**cetabulum **D**iameter	With the bone in anatomical position, measure from the inner border closest to the middle of the ischium to the superior border in line with the notch below the anterior iliac spine.	(Bubalo, 2019)
TAD	**T**ransverse **A**cetabulum **D**iameter	With the bone in anatomical position measure from the inner border of the joint surface closest to the superior border of the obturator foramen to the border in line with the sciatic notch.	(Bubalo, 2019)
FHD	**F**emur vertical **H**ead **D**iameter	Maximum diameter of femoral head in the vertical plane taken from the border of the articular surface.	(Asala, 2001)
THD	Femur **T**ransverse **H**ead **D**iameter	Maximum diameter of the femoral head taken in the horizontal plane perpendicular to FHD.	(Asala, 2001)
MCL	Femur **M**edial **C**ondylar **L**ength	Distance between most anterior and most posterior projection of the medial condyle.	(Asala, 2004)
BB	Femur **B**icondylar **B**readth	Distance between the two most laterally projecting points on the epicondyles. Measurement parallel to the distal surfaces of the condyles.	(Asala, 2004)
LCL	Femur **L**ateral **C**ondylar **L**ength	Distance between most anterior and most posterior projection of the lateral condyle.	(Asala, 2004)
TPB	**T**ibia **P**roximal **B**readth	Maximum distance between the two most laterally projecting points on the medial and lateral condyles of the proximal epiphysis.	(Steyn & Işcan, 1997)
GL	**G**lenoid prominence **L**ength	Measured from the superior margin to the inferior margin on the glenoid prominence.	(MacAluso, 2011)
GB	**G**lenoid prominence **B**readth	Distance across glenoid cavity measured perpendicular to GL.	(MacAluso, 2011)
EB	Humerus **E**picondylar **B**readth	Distance from most laterally protruding point on the lateral epicondyle to the corresponding point on the medial epicondyle.	(Steyn & Işcan, 1999)
HHD	**H**umerus Vertical **H**ead **D**iameter	Maximum diameter of the head of the humerus taken in the vertical plane from the border of the articular surface.	(Steyn & Işcan, 1999)

4.2 STATISTICAL ANALYSES

BASELINE STATISTICAL TESTING

Statistical analyses were performed in MATLAB_R2020b 9.9.0.1495850 under an Academic licence. A 95% significance level and a p-value of 0.05 or less were considered significant. All MATLAB scripts can be found at: https://github.com/mscott1037/masters.

4.2.1 OBSERVER AGREEMENT

A subset of individuals was randomly selected and resampled to assess observer agreement. This included 30 individuals for inter-observer agreement and 40 individuals (20 resampled after data collection concluded at the UCT skeletal repository and 20 resampled after data collection concluded at the Dart collection) for intra-observer agreement. The individuals were selected using a feature built into the custom PostgreSQL database which allows the user to randomly select a skeleton number which already exists in the database.

All the parameters measured in this study were continuous numerical variables. One of the many statistical tests which can be used to assess rater reliability for continuous numerical data is Lin's Concordance Correlation Coefficient (CCC).

Pearson’s *r*, a different measure of association, is commonly used to assess correlation between 2 variables. However, CCC tests for both correspondence and covariation which Pearson’s *r* does not account for. This means that where Pearson would still give a perfect score when the data are shifted up or down and the data trend line does not pass through the origin, CCC takes these factors into account and would provide a more accurate correlation estimate (Lin, 1992).

CCC was used to assess the degree of agreement between the new measurement Y and the initial observation X (Lin, 1992) The function *f_CCC* (Matthew, 2021) was used to perform the analysis. The output, CCC, was a value between 0 and 1 with 0 signifying no correlation and 1 being perfect correlation. A score between 0.9 and 0.99 was considered acceptable.

In addition to CCC testing, a scatter plot was charted for the mensuration where the CCC value was lowest.

4.2.2 OUTLIERS

Outliers - data points that differ markedly from others in a dataset - can be defined, identified, and handled in a multitude of ways. Outliers can be separated into 3 main categories: error outliers, interesting outliers and influencing outliers (Aguinis, Gottfredson & Joo, 2013).

Error outliers are nonlegitimate observations resulting from errors measuring, recording and/or preparing the data. These outliers can be identified in many visual ways including boxplots, stem and leaf diagrams and Q-Q plots or using numeric methods like percentage analysis or Mahalanobis distance to name a few (Aguinis, Gottfredson & Joo, 2013). During data collection, each mensuration was measured 3 times for each observation to reduce potential error. The benefit of this technique was the ability to check for agreement between the trio of measurements for each observation. The most efficient way to discover these outliers in the study data was to run all the data through a filter which highlighted any observations where the difference in measurement value for repeated measurements exceeded 1mm. These entries were then individually examined to determine if an error outlier was present.

Error outliers can be handled in a number of ways. Most often it is recommended that error outliers are corrected where possible, remeasured if there is uncertainty or, removed – when remeasurement is not possible and the error cannot be corrected (Jackson & Chen, 2004; Aguinis, Gottfredson & Joo, 2013). Obvious errors (like a typing error where 2 measurements were very similar, but the 3rd had a substitution, or a missing/ misplaced decimal point) were handled by correction. Inconclusive entries were remeasured where possible or removed where an error was obvious but remeasurement was not viable (if 1 of the 3 measurements was possibly erroneous then it was removed, and the other 2 measurements were retained).

Error outliers were thus defined as mistakes in the data, identified using a filter and then corrected or removed.

Interesting outliers are datapoints that differ markedly from others but cannot be credited to error. Influencing outliers are those interesting outliers which have a significant impact on the structure of the dataset (Aguinis, Gottfredson & Joo, 2013). Interesting and influencing outliers represent the fringes of possibility in a population or group and were detected using a boxplot (MathWorks, 2006a) and defined as any observation more than 1.5 times the interquartile range (IQR) away from the top or bottom edge of the IQR. These values were denoted by a red cross in the boxplots.

4.2.3 NORMALITY TESTING

A normal distribution (also known as a Gaussian Distribution) is a probability distribution that is symmetrical about the mean with data near the mean being more frequent than data further from the mean (Mishra et al., 2019). An underlying normal distribution in the data is often an assumption for statistical tests.

Population vs. Sample

When evaluating a population, it is generally not possible to collect data from the entire population. The *population* in this study refers to every single male and female South African of SAAA, SAMA and SAEA, a total of 57.25 million people (Statistics South Africa, 2019). The study *sample* is thus a representation of the population.

Population *parameters* include means, standard deviations, and other proportions, however when the full population is unattainable and only a sample is available, we cannot know these parameters, so we calculate *statistics* from our sample data to estimate the true values.

Parametric vs. Non-parametric

Parametric statistical techniques rely on assumptions about the underlying population such as the shape and parameters of the assumed distribution. When these assumptions are not met, parametric procedures are unreliable, and it is inaccurate to apply them. Conversely, nonparametric statistical techniques tend to rely on few to no assumptions about the shape and parameters of the population distribution from which the sample was drawn.

Many of the more commonly used statistical tests such as t-tests and ANOVA rely on the population having a normal distribution. We do not know the real *population* parameters for any of the mensurations in this study, so normality testing procedures needed to be implemented.

Since parametric tests are so reliant on underlying qualities of the population, why don't we always use non-parametric alternatives? Two main drawbacks exist. The first weakness of nonparametric methods is that they are statistically less powerful than the analogous parametric method when the data are truly approximately normal (Mishra et al., 2019). Less statistical power means that the probability of estimating an association between two truly associated variables is smaller. To counteract this, a larger sample size would be needed to have the same power as the corresponding parametric test (Hoskin, 2012). The second drawback is ease of interpretation. Nonparametric procedures tend to have more complex results or rely on medians rather than means.

Non-parametric procedures can be a useful alternative in many cases but are not always the most accurate and reliable solution. Normality testing is thus a vital step to ensure the most powerful testing procedure is followed for the dataset in question.

Methods for Testing Normality

To test for normality in a sample data (for continuous variables), subjective visual methods (histograms/Q-Q plots) or objective, but often overly sensitive formal, numerical normality tests can be employed (Mishra et al., 2019). Statistical normality tests include Shapiro-Wilk (Shapiro & Wilk, 1965; Shapiro & Francia, 1972), Kolmogorov-Smirnov (Massey, 1951; Miller, 1956; Marsaglia, Tsang & Wang, 2003; Steinskog, Tjøtheim & Kvamstø, 2007), Lilliefors (Lilliefors, 1967, 1969), Anderson-Darling, Jarque-Bera (Thadewald & Büning, 2007) and a host of other options (Arnastauskaitė, Ruzgas & Bražėnas, 2021).

The Kolmogorov-Smirnov test (kstest) for normality is a valid statistical method for both large and small sample sizes, although noted to be overly sensitive on samples <50 (Mohd Razali & Bee Wah, 2011). The *kstest* is an empirical cumulative distribution function, so goodness-of-fit is determined by comparing the data to an empirical distribution function (Massey, 1951). The drawback to this methodology is that it is highly sensitive to extreme values and is therefore likely to reject normality (Marsaglia, Tsang & Wang, 2003).

This overly conservative nature of the kstest is 'corrected' in the Lilliefors test (Lilliefors, 1967). The Lilliefors test uses a Monte Carlo calculation to correct for extreme values and makes the kstest less conservative. In the case of this study, being conservative is not a negative trait as the intention is to exercise caution in applying parametric testing procedures.

Due to its conservative nature, although differently employed by different software packages, the kstest been noted to be unreliable in some cases with the recommendation to rather apply a Shapiro-Wilk test or a Jarque-Bera test (Steinskog, Tjøtheim & Kvamstø, 2007).

The *swtest* is regularly used in the biological sciences as a normality test but was not selected as a normality testing procedure for this study as it is not supported by MATLAB.

The *jbtest* test can be used in addition to the *kstest* as an added stringency measure. The *jbtest* is a goodness-of-fit test which assesses whether the sample data have skewness and kurtosis (Mishra et al., 2019) matching that of a standard normal distribution. A true normal distribution has a skewness of 0 (data are symmetrical) and a kurtosis of 0 (inverted parabolic shape). Thadewald & Büning (2007) claim that Jarque-Bera testing is superior on datasets with medium to long tails and a symmetric distribution, however it has lower power than some of its competitors when applied to a sample which has short tails or is bimodal. It must be noted that this study sample, which contains 2 classes (male and female) is likely to be bimodal (and thus not reflect a standard normal distribution) and may thus follow a gaussian mixture distribution. Despite concerns raised by Thadewald & Büning (2007), the jbtest is a well-recognised, rigorous statistical test which is endorsed by many statisticians (Arnastauskaitė, Ruzgas & Bražėnas, 2021).

To test for normality, a one-sample Kolmogorov Smirnov test (*kstest*) and a Jarque-Bera (*jbtest*) were performed on each of the left- and right-side specific measurements (Massey, 1951; Miller, 1956; Marsaglia, Tsang & Wang, 2003; Thadewald & Büning, 2007).

The *kstest* (MathWorks, 2021a) and *jbtest* (MathWorks, 2021b) are hypothesis tests which examine the null hypothesis (H_0) that the data follow a standard normal distribution with the alternate hypothesis that the data are not normally distributed. H_0 is rejected when $p<0.05$ and the alternative hypothesis is rejected when $p>0.05$. A small p-value ~0.05 would cast doubt on the validity of the null hypothesis while a larger p-value would provide evidence to suggest that the null should not be rejected.

DEMOGRAPHIC ANALYSIS

Given the aim to use the sample data to estimate sex, it is vital to investigate potential confounding factors which may impede sex estimation. Additionally, it is also important to determine whether the mensurations are sexually dimorphic and able to discriminate sex.

4.2.4 BILATERAL ASYMMETRY

Bilateral asymmetry, the size- and shape -based differences between the left- and right- sided bones of the skeleton, is an important consideration when performing skeletal analysis. Do we use left bones only? Right bones only? A combination of both?

Testing for Bilateral Asymmetry

Bilateral variation is commonly assessed using a paired sample t-test (Dabbs & Moore-Jansen, 2010; Dillon, 2014; Carvallo & Retamal, 2020; Jerković et al., 2020) however, these are parametric tests. Therefore, a 2 sample Wilcoxon Signed Rank test (*signrank*) has been applied for each mensuration.

The *signrank* test (MathWorks, 2021c) is a nonparametric hypothesis test which examines the null hypothesis that paired values (in this case: left- and right- sided measurement from the same individual) have a median difference of zero against the alternate hypothesis that the median difference is not zero (the left- and right-sided measurements differ significantly). As with all other hypothesis tests if p<0.05 then the null hypothesis is rejected.

Median values for left and right measurements as well as standard error (SE) and percentage of pairs with:

- left > right
- left < right and
- left = right

were calculated for each mensuration.

A variation on a violin plot (Jonas, 2008) was charted for each mensuration to visualise the distribution of values for left- and right-sided measurements. Violin plots are like box plots but with a rotated kernel density plot on either side of a y-axis. They visualise the probability density of the data at different values. In this case, a variation of a violin plot has been used where, instead of a reflection of all the data about the y-axis, the male data is reflected whilst the female data remains on the other side. This allows one to assess if there are any marked differences in bilateral patterns between males and females.

4.2.5 SEXUAL DIMORPHISM METHODS

The characteristic differences between male and female skeletons, defined as sexual dimorphism, are the most effective attributes with which to assess skeletal sex. The more dimorphic any particular trait is, the more effective it should be in classifying sex.

As previously mentioned, males and females are mostly androgynous until puberty when they begin to sexually diverge. The main contributor to sexual divergence is sex steroid hormones during pubertal development (Wells, 2007). For example, oestrogen is important in the pattern of female bone development but unfortunately also predisposes females to a greater risk of osteoporosis in old age (Wells, 2007). A host of other complex factors and processes contribute to sexual dimorphism in both bone size and shape including inequality in food provision during development (Akachi & Canning, 2007), genetics (Rawlik, Canela-Xandri &

Tenesa, 2016), ICT (Hochberg et al., 2011) and adaptation to parturition (Stulp & Barrett, 2016).

It is important to assess whether mensurations are sexually dimorphic as if they are not, then they should be ruled out as potential predictors.

Sexual dimorphism was assessed using a non-parametric hypothesis test called a Wilcoxon Rank Sum Test (*ranksum*). The *ranksum* test (MathWorks, 2021d) compares the data in 2 sample to determine whether the data from both follow continuous distributions with equal medians. The null hypothesis is that the data do follow continuous distributions with equal medians whilst the alternate hypothesis is that they do not. This test returns a p-value which provides "the probability of observing a test statistic as extreme or more extreme than observed under the null hypothesis" (MathWorks, 2021d). The smaller the p-value, the less likely the null hypothesis is. As with all other hypothesis tests if $p<0.05$ then the null hypothesis is rejected.

4.2.6 SECULAR TREND AND AGE-AT-DEATH

Using skeletal collections as proxies for the South African population bears the risk of presenting a sampling bias towards individuals of certain age-groups. The UCT cadaveric skeletal repository, for example, is composed of mostly older, SAEA individuals (Maass & Friedling, 2019). In this case, risk lies in not only overrepresenting a particular age-group but also with the potential for a disparity between the age-ancestry group demographics.

Correlation between year-of-birth and measurement value was assessed using Pearson's Correlation Coefficient (*r*). Correlation was calculated using the *corr* function (MathWorks, 2006b) which returns the correlation coefficient and a p-value which tests the hypothesis of no correlation against the alternate hypothesis of a nonzero correlation. The value of *r* ranges from -1 to +1. A perfect negative linear correlation would have a value of -1, no correlation would have a value of 0 and a perfect positive correlation would have a value of +1.

Age-at-death in the study sample compared to UCT and Dart skeletal repositories was plotted in a series of histograms. A bar chart including age-at-death data for the South African population (Statistics South Africa, 2018) was also charted along with a histogram depicting year-of-birth of individuals in the study sample. Demographic details for the study sample including age and year-of-birth were also summarised.

Paired sample linear correlation (Pearson's *r*) was used to investigate the relationship between age-at-death and measurement value for each mensuration.

4.3 MACHINE LEARNING MODELLING

Data mining, or knowledge discovery, is the multistep workflow typically associated with solving a machine learning problem. The goal here was to classify data into classes. The first step was to pre-process the data into a manageable and useful format. The processed data was then used to train and test sex classification models and a do a sample size analysis.

STEP 1: DATA PREPROCESSING

Prior to model training, data were pre-processed to select the most relevant features, ensure reliability, partition data, remove incomplete data and identify predictors.

4.3.1 FEATURE SELECTION

Feature selection aims to improve prediction performance of ML models and provide quicker, more cost-effective classifiers by reducing the dimensionality of data (Iguyon & Elisseeff, 2003). This is achieved by selecting only a subset of the predictor variables, when training a model, that best model the desired response. Using too many features in a prediction model leads to overfitting and can degrade prediction potential even when all the features included are relevant and contain information about the response variable.

All predictors in the dataset were ranked and predictor importance scores were calculated using the *fscmrmr* function (MathWorks, 2019). The MRMR (minimum redundancy maximum relevance) algorithm is used to rank features (predictors) sequentially. Predictor importance scores for the 12 variables were plotted on a bar graph.

4.3.2 REPLICABLE AND REPRODUCIBLE

In applied machine learning, a learning *algorithm* is run on a training dataset to produce a machine learning *model*. The *model,* which contains the data structure and coefficients required to make predictions, is then validated using data not used during training or applied to new data to make predictions. This workflow is communicated in Figure 4-2.

Reproducibility and replicability are important and intensely debated topics in the field of science and medicine (Beam, Manrai & Ghassemi, 2020). With the retraction of many papers due to unreproducible results, it Is believed that science has found itself in a reproducibility crisis (McNutt, 2014). A survey published by *Nature* found that 34% of respondent scientists had not established procedures for reproducibility in their labs (Baker & Penny, 2016). Reproduction and validation of scientific findings across different research groups and study populations and is essential and forms the basis for why the scientific method is followed when recording and communicating one's research outputs.

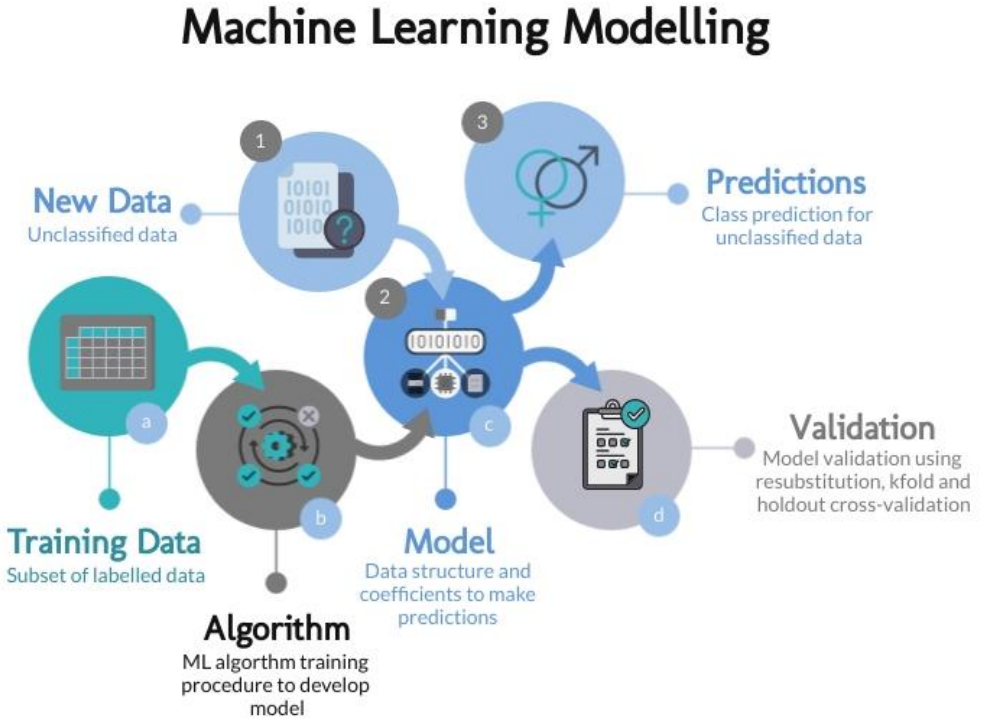

Machine Learning Modelling
1
New Data
Unclassified data
3
Predictions
Class prediction for unclassified data
2
10101010
a
b
c
d
Training Data
Subset of labelled data
Algorithm
ML algorthm training procedure to develop model
Model
Data structure and coefficients to make predictions
Validation
Model validation using resubstitution, kfold and holdout cross-validation

Figure 4-2 Machine learning workflow for algorithm training, validation and use on new data.

By definition, a study is *reproducible* if given access to the underlying data and analysis code, an independent group can obtain the same results observed in the original study. Saying that a study is reproducible doesn't imply that it is correct, only that the results could be verified. A study is *replicable* if an independent group can reach the same conclusion after repeating the same set of experiments or analyses on new data (Held & Schwab, 2020).

Earlier in this study, for example, inter- and intra-observer error were assessed. Ensuring that repeated measurements by the same observer and an independent observer are in agreement aides in ensuring that the sampling portion of the study is *replicable*.

In machine learning, function approximation is the process of 'learning' or searching for a function which maps inputs to outputs. There are usually many possible ways for the inputs to be mapped to the outputs and thus small differences in the training algorithm, validation process or dataset can result in a different model, level of error or, predictions when evaluating new data.

Many strategies exist to ensure that the best possible path between inputs and outputs is chosen to minimise differences and ensure replicability and reproducibility. An in-depth explanation of strategies to ensure reproducibility and replicability and, reduce variance within this study can be found in Appendix B.

As a standard practice, a step was added to the beginning of all scripts to set the seed for the random number generator (rng). Always using the same seed helps to ensure reproducibility of modelling.

4.3.3 PARTITIONING OF DATA

The data were partitioned into training and testing subsets using the *cvpartition* function (MathWorks, 2008). The *cvparition* function was used to ensure than an equal number of males and females were randomly selected and removed from the training subset of the data. This 'test' subset of 20% was not involved in algorithm training but rather retained and used after training was complete to perform a final validation of results.

4.3.4 MISSING DATA

As is the nature of skeletal data, entries for individuals are often incomplete. In many cases. bones were missing, damaged or exhibited pathologies or trauma and were thus not sampled. This led to many observations being 'incomplete'. Before each algorithm was trained, relevant data were selected and then all incomplete entries were removed from both the training and testing subsets. Removal of incomplete data was performed after predictor selection and data partitioning to maximise the number of individuals in the dataset. Sample sizes for each model were calculated and tabulated.

4.3.5 TRAINING VARIABLES

The following combinations of predictors were used in algorithm training:

1. Each of the 12 predictor variables for univariate models
2. Bone models:
   a. Pelvis (AD, TAD)
   b. Femur (FHD, THD, MCL, BB, LCL)
   c. Tibia (TPB)
   d. Scapula (GL, GB)
   e. Humerus (EB, HHD).
3. Joint models:
   a. Hip (AD, TAD, FHD, THD)
   b. Knee (MCL, BB, LCL, THD)
   c. Shoulder (GL, GB, HHD).
4. A multivariate model using all 12 predictor variables.
5. An optimised model using the most relevant predictors and least redundant number of predictors.

4.3.6 SAMPLE SIZE

Data were partitioned into training and testing subsets and processed to remove incomplete observations. The number of observations in the training and testing subsets for each model is summarised in Table 4-2. The all-predictors model (AP) contained the least observations as only complete skeletons could be included in this dataset.

Table 4-2 Sample size and predictors for classification models.

		Training subset			**Testing subset**		
Model	Predictors	Subset	Male	Female	Subset	Male	Female
		[520]	[261]	[259]	[130]	[65]	[65]
Univariate							
AD	AD	506	249	257	123	62	61
TAD	TAD	506	249	257	123	62	61
FHD	FHD	489	240	249	119	60	59
THD	THD	489	240	249	119	60	59
MCL	MCL	489	240	249	119	60	59
BB	BB	489	240	249	119	60	59
LCL	LCL	489	240	249	119	60	59
TPB	TPB	452	217	235	112	57	55
GL	GL	494	244	250	120	60	60
GB	GB	494	244	250	120	60	60
EB	EB	489	237	252	122	61	61
HHD	HHD	489	237	252	122	61	61
Bone							
Pelvis	AD/TAD	506	249	257	123	62	61
Femur	FHD/THD/MCL/LCL/BB	489	240	249	119	60	59
Tibia	TPB	452	217	235	112	57	55
Scapula	GL/GB	494	244	250	120	60	60
Humerus	EB, HHD	489	237	252	122	61	61
Joint							
Hip	AD/TAD/FHD/THD	483	235	248	118	60	58
Knee	MCL/BB/ LCL/TPB	488	239	249	110	56	54
Shoulder	GL/GB/EB	480	233	247	120	60	60
All	All*	415	193	222	104	53	51
Optimal	LCL/TPB/GL	425	203	222	105	53	52

*AD/TAD/FHD/THD/MCL/BB/LCL/TPB/GL/GB/EB/HHD
[n total]

STEP 2: TRAINING SEX CLASSIFIERS

Machine learning (ML) is a subdivision of artificial intelligence (AI) within the field of computer science which uses data and algorithms to 'learn' by gradually improving accuracy to predict, or respond to, future data (Sen, Hajra & Ghosh, 2020). ML has many applications including: facial recognition, speech-to-text, path planning and targeted advertising; where it is impractical, or even impossible, to manually write algorithms to complete the task.

Historically, the major difference between humans and computers has been that humans innately endeavour to improve their way of tackling a problem when they fail or achieve sub-par results whilst computers haven't looked at their results and have thus been unable to automatically improve their approach to completing a task without being explicitly programmed to do so (Kidwell, 2015). In response to this problem, the field of ML evolved.

ML involves the generation of computer programs that can automatically learn and improve their performance using assemblages of data and learned experience.

4.3.7 ALGORITHM TRAINING

Just as people need to be taught to execute tasks, ML systems require training. Programmers achieve this by providing the system with inputs, and corresponding outputs, and adjusting the structure of the model, or data in the learning machine, so that inputs are mapped to the correct outputs (Tamir, 2020). If enough training pairs exist, then the machine should be able to select the correct outputs when new inputs are presented; but if the training data are insufficient and fail to accurately represent the full range possibilities then problems are likely to arise (Paluszek & Thomas, 2017).

Training can be subdivided into supervised and unsupervised learning. *Unsupervised* learning is applied to problems where there isn't necessarily a 'right' answer. The most simplified explanation is that data in unsupervised learning are not *labelled*. The general aim of unsupervised learning is to cluster data points based on similarities (Sommer & Gerlich, 2013). Unsupervised learning can be used, for example, to disseminate between left- and right- sided bones given a series of photographs with no attribute data. Conversely, *supervised* learning involves the use of specific training sets of data which are *labelled.* In essence, each input is expressly mapped to a specific output (Alloghani et al., 2020). An example of this would be training an algorithm to predict whether a photographed bone is human, given a training dataset of many photographs of bones labelled as human or non-human.

Typically, a supervised ML algorithm consists of 3 modules (Tamir, 2020). The first is a decision process, which can be thought of as the 'recipe', or a series of calculations or steps that transform the input data into a prediction. This decision process contains a set of weights which relate each input to the desired output (or set of outputs). The second module, an error function, measures the goodness of fit of the model by comparing the predicted output to known outputs. This module asks: "Did the decision process make the correct prediction?". The final module involves analysing where the decision process went wrong, using the output of the error function, and then updates the weights in the decision process accordingly to attempt to minimize the error function.

An optimisation algorithm will repeat these steps and subsequently optimize the decision process module iteratively, updating the weights in this module until a threshold accuracy, or local minimum (relative to some set of starting weights), is met (Tamir, 2020).

There are two types of supervised learning: regression and classification. Regression problems output continuous data whereas the outputs of classification problems are categorical (Sen, Hajra & Ghosh, 2020).

Sex estimation, as a machine learning problem, is a supervised classification problem. Given a set of labelled input measurements, the goal is to predict sex - a categorical variable.

4.3.8 ML ALGORITHMS

Whilst many ML algorithms exist with a multitude of use cases, the algorithms best suited to this classification problem are *k*-Nearest Neighbour, decision trees, discriminant function analysis, Naïve Bayes, and a series of ensemble methods. An overview of these algorithms will be presented in the sections to follow. For each variable or combination of variables in section 4.3.5, the following algorithms were trained and hyperparameters were auto optimised (unless otherwise stated).

4.3.8.1 NAÏVE BAYESIAN NETWORKS (NB)

Naïve Bayes (NB) classifiers are built on the concept of conditional probability as described by the Bayes Theorem, which first appeared in the 18th century (Berrar, 2018). Bayes theorem, seen below, provides the basis for calculating posterior probability.

$$\underset{\text{Posterior probability}}{p(c|x)} = \frac{\overset{\text{Likelihood}}{p(x|c)} \times \overset{\text{Class prior probability}}{p(c)}}{\underset{\text{Predictor prior probability}}{p(x)}}$$

$$p(c|X) = p(x_1c) \times p(x_2c) \times \ldots p(x_nc) \times p(c)$$

$p(c|x)$ posterior probability of class (c) given predictor variable x

$p(x|c)$ probability of predictor (x) given class (c)

$p(c)$ probability of class c

$p(x)$ prior probability of predictor (x) occurring

The NB algorithm calculates the conditional probability of an outcome occurring, based on prior knowledge of conditions that might be related to the event (Alloghani et al., 2020). The algorithm produces a frequency matrix from the dataset and then uses this to produce a marginal probability matrix by finding the probabilities of events or combinations of events occurring. When new data or test data is applied to the model, the Naïve Bayesian equation is then used to calculate the posterior probability for each class given a set of attributes. The class with the highest posterior probability is then chosen. A simplified visual summary of Naïve Bayes classification is provided in Figure 4-3.

Although much simpler than more sophisticated learning algorithms like artificial neural networks, NB classifiers have proven superior even on datasets with substantial feature dependencies and smaller sample sizes (Osisanwo et al., 2017).

NB generally functions on the somewhat unrealistic assumption of independence between predictor variables however, MATLAB documentation claims that the algorithm "still appears to work well in practice" even when this assumption is not met (MathWorks, 2021e). Another default assumption made by the NB algorithm is that the data follow a normal distribution (MathWorks, 2021e). Given that the data in this study do not fit a normal distribution, the *fitmethis* function (de Castro, 2021) was used to determine what the best fit distribution is for the data. The best fitting distribution was found to be a kernel distribution.

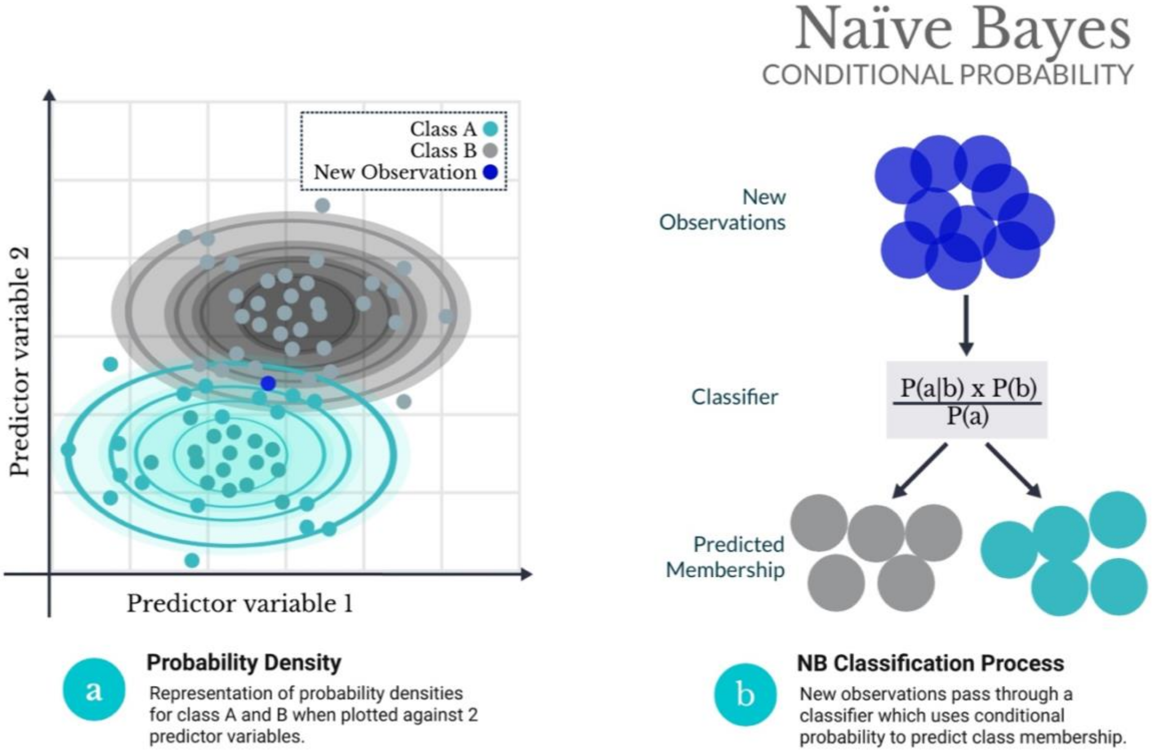

Naïve Bayes
CONDITIONAL PROBABILITY
Class A ●
Class B ●
New Observation ●
Predictor variable 2
Predictor variable 1
New Observations
Classifier
P(a|b) x P(b) / P(a)
Predicted Membership
a
Probability Density
Representation of probability densities for class A and B when plotted against 2 predictor variables.
b
NB Classification Process
New observations pass through a classifier which uses conditional probability to predict class membership.

Figure 4-3 Visual explanation of a Naive Bayes Classifier.

A kernel distribution is a non-parametric representation of a variable's probability density function (pdf) and is used when a parametric (normal) distribution is not an accurate elucidation of the data or when one wants to avoid making assumptions about the data (Bowman & Azzalini, 1999). Hyperparameters which help to define a kernel distribution include a smoothing function and a bandwidth value which control the parameters of the resulting density curve.

The calculated estimated pdf of any random variable is known as a kernel density estimator. The formula for a kernel density estimator is as follows:

$$\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

For any real (continuous numerical) values of x, where $x_1 \rightarrow x_n$ are random samples, n represents the sample size, $K(\dots)$ the smoothing function, and h denotes the bandwidth (MathWorks, 2013).

The *fitcnb* function (MathWorks, 2014a) was used to train a NB algorithm for each predictor variable or set of predictor variables. The distribution name was specified using the [Name, Value] pair: 'DistributionNames', 'kernel'. Bandwidth and smoothing functions were optimised for as part of the algorithm training process.

4.3.8.2 K-NEAREST NEIGHBOURS (KNN)

K-Nearest Neighbours (KNN) classifies new data by finding the nearest neighbouring datapoints to a new query by searching an existing database of observations (Cunningham & Delany, 2021). This is illustrated in Figure 4-4 where the dark blue circles represent the 'search area' for *k* similar data points. The KNN algorithm is supposedly competitive with even the most precise models as it is said to have the ability to make highly accurate predicitions (Sen, Hajra & Ghosh, 2020).

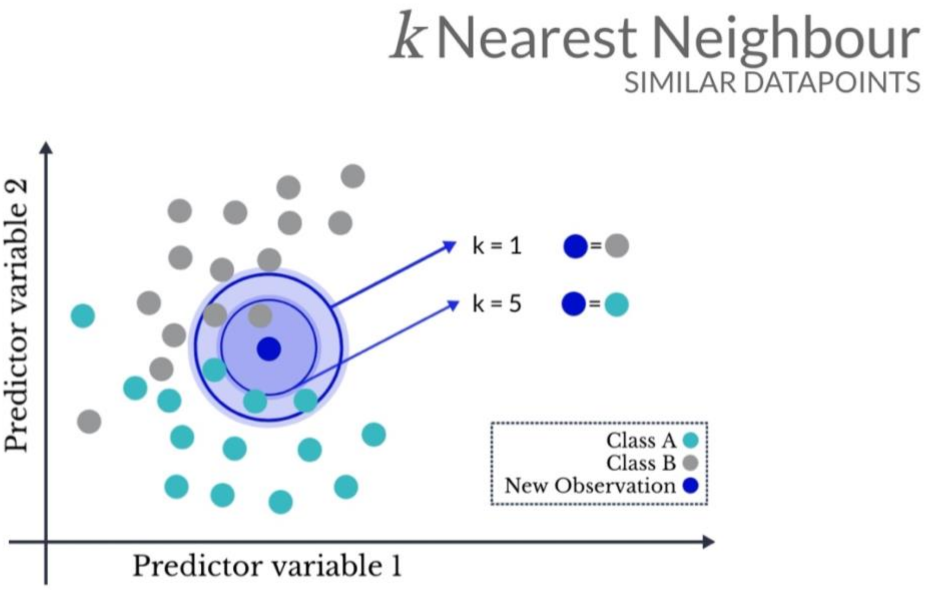


Figure 4-4 Representation of K-Nearest Neighbour classification with 2 predictor variables. Changing the value of k from 1 to 5 causes a different prediction outcome.

Distances between datapoints can be calculated in several ways including Cityblock, Minkowski, Mahalanobis and Euclidean distances (Cunningham & Delany, 2021). When more than 1 predictor is used, distance will need to be calculated in multiple dimensions. Euclidean distance will be used as an example to explain how this is done. Euclidean distance between 2 points is calculated in 1 dimension using the equation below:

$$Distance = \sqrt{(x_A - x_B)^2}.$$

The descriptor variable is specified by x with x_A being the new datapoint and x_B being a reference datapoint. This function can be applied in multiple dimensions by simply adding descriptor variables or dimensions:

$$Distance = \sqrt{(x_A - x_B)^2 + (y_A - y_B)^2 + (z_A - z_B)^2 \dots}$$

The distance between the new unknown datapoint and all existing datapoints is calculated and the nearest datapoints are selected. A voting system is then used to calculate the predicted class of the new datapoint.

Several training parameters can be specified or optimised for. Amongst optimisable hyperparameters, weightings can be applied so that nearer datapoints have ‘more voting power’ than those further away, different ways to calculate distance can be specified and the optimal value of *k* – the number of nearest datapoints to assess – must be determined.

The value of *k* is always an important consideration when using KNN as the value of *k* may change prediction outcomes and thus the accuracy of the model as shown in Figure 4-4. If the value of *k* is too large then the intention of KNN to predict class based on proximity to an observation’s neighbours is lost but if *k* is too small, it will not account for sufficient variation (Guo et al., 2003).

Due to its nature as a “lazy learning” algorithm, where the computation for predictions is deferred until classification of new instances, KNN has high computational costs (Guo et al., 2003). These increased costs can be overlooked where predictions are not requested frequently but high accuracy is required. Sex estimation from skeletal remains, where a single individual needs to be identified with accuracy paramount, is thus the perfect application for this type of learning model. Conversely, KNN would be inefficient if one were to attempt to classify thousands of new observations in a short amount of time.

The *fitcknn* function (MathWorks, 2014b) was used to train models for each predictor variable or set of predictor variables. All hyperparameters were auto optimised using the built-in optimisation argument.

4.3.8.3 DECISION TREES

Binary decision trees (Breiman et al., 1984) are logical learning algorithms which use a series of if-then rules to make a prediction of class. When fitting a tree to continuous numerical data, the model will determine a sectioning point for each variable by iteratively splitting the dataset on the dependent variable that best separates the data into the different existing classes (Kotsiantis, 2007). The general structure of a binary decision tree is described in Figure 4-5.

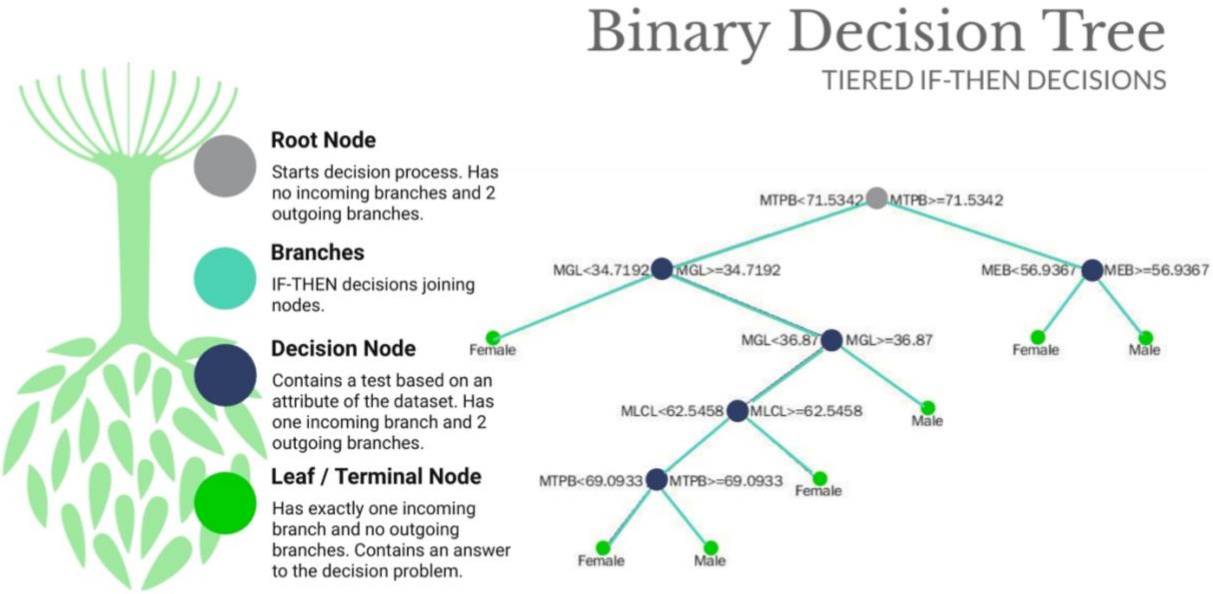

Binary Decision Tree
TIERED IF-THEN DECISIONS
Root Node
Starts decision process. Has no incoming branches and 2 outgoing branches.
Branches
IF-THEN decisions joining nodes.
Decision Node
Contains a test based on an attribute of the dataset. Has one incoming branch and 2 outgoing branches.
Leaf / Terminal Node
Has exactly one incoming branch and no outgoing branches. Contains an answer to the decision problem.
MTPB<71.5342
MTPB>=71.5342
MGL<34.7192
MGL>=34.7192
MEB<56.9367
MEB>=56.9367
Female
MGL<36.87
MGL>=36.87
Female
Male
MLCL<62.5458
MLCL>=62.5458
Male
MTPB<69.0933
MTPB>=69.0933
Female
Female
Male

Figure 4-5 Description of the structure of a binary decision tree with a sex classifier as an example.

Fitctree (MathWorks, 2014c) performs various statistical tests to compute variable association, predictor-response interaction, and impurity (fraction of observations from the 2 classes on either side of a sectioning point) to assist in node splitting. Weights are then assigned to observations based on purity and association. A thorough mathematical description of node splitting within the algorithm can be found at MathWorks (2014c).

An important part of multivariate tree training is pruning. Pruning is a data compression technique in ML which reduces the size of decision trees by removing non-critical and redundant portions of the tree (Osisanwo et al., 2017). This pruning process means that even with added variables, 2 trees might be trained and optimised to have the same number of splits because the added variables are redundant and do not significantly improve the model (Dietterich & Kong, 1995).

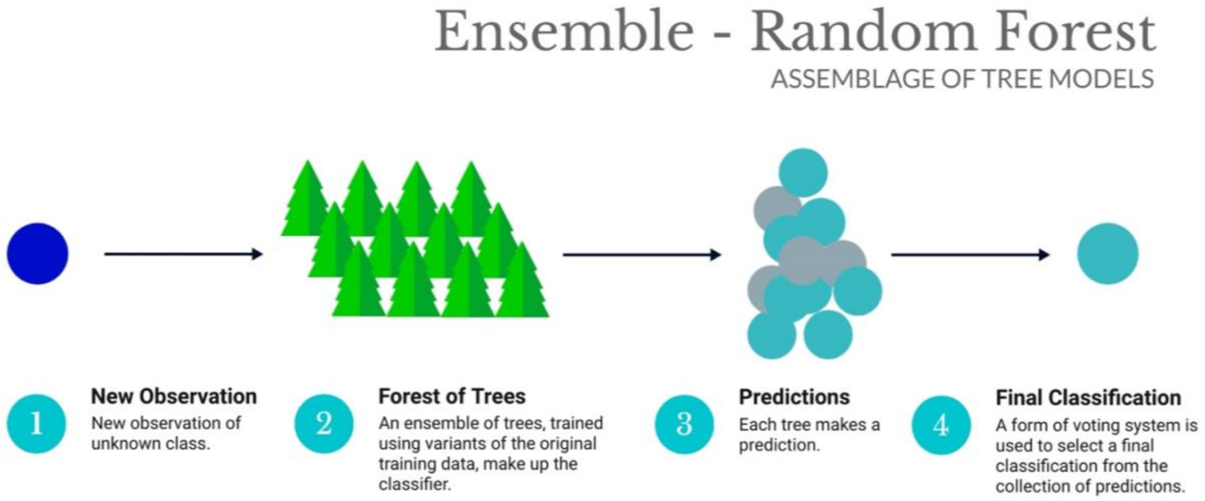
The *fitctree* function was used to train models for each predictor variable or set of predictor variables. Hyperparameters were optimised. Controlling hyperparameters can aid in avoiding overfitting of a tree model. For example: setting the min leaf size ensures that there are a minimum number of instances within each terminal node. Sectioning points were extracted from each model output and where needed; the tree viewer was used to view the structure of trained decision trees.

4.3.8.4 ENSEMBLE METHODS

Ensemble classifiers are conglomerates of bagged, boosted or subspace sampled classification models (MathWorks, 2018). Several base models are combined to create a single optimal predictive model. Base models can include Naïve Bayes, KNN, discriminant functions, and decision trees.

Some ML methods are considered 'weak' learners as they are highly sensitive to the training data (MathWorks, 2018). A small change to the training data can result in 2 completely different models and consequently, 2 different predictions. But which prediction is correct?

Ensemble methods exploit the weaknesses of learners, like decision trees, by growing an *ensemble* of trees, analogously referred to as a forest. This is illustrated in Figure 4-6.



Ensemble - Random Forest
ASSEMBLAGE OF TREE MODELS
1
New Observation
New observation of unknown class.
2
Forest of Trees
An ensemble of trees, trained using variants of the original training data, make up the classifier.
3
Predictions
Each tree makes a prediction.
4
Final Classification
A form of voting system is used to select a final classification from the collection of predictions.

Figure 4-6 Diagram representing the workflow of an ensemble classifier.

New observations are classified using the assemblage of models, in this example - a random forest, instead of just a single classifier (Ren, Zhang & Suganthan, 2016). The predictions of multiple models are then combined using a form of voting system to select the final prediction (Figure 4-6). The premise behind ensemble is to aggregate many weighted models to obtain a combined model that bests all the models in it (Ren, Zhang & Suganthan, 2016).

4.3.8.4.1 BAGGING

An example of an ensemble learning method is **B**ootstrap **Agg**regation (bagging). Bagging is used to generate many bootstrap replicas of the dataset and then grow an ensemble of classifiers on the replicas (González et al., 2020). Each classifier will be slightly different, and an averaging or majority vote system is used to combine the learners.

4.3.8.4.2 BOOSTING

In comparison, boosting algorithms train sequential models, each of which attempts to correct the errors of the previous model until an optimal model is reached (González et al., 2020). In Figure 4-7, one can see how bagging and boosting differ. When using boosting, the original data is used to train a model which feeds back to the data and weights it. This process can be repeated many times until the final model is selected.

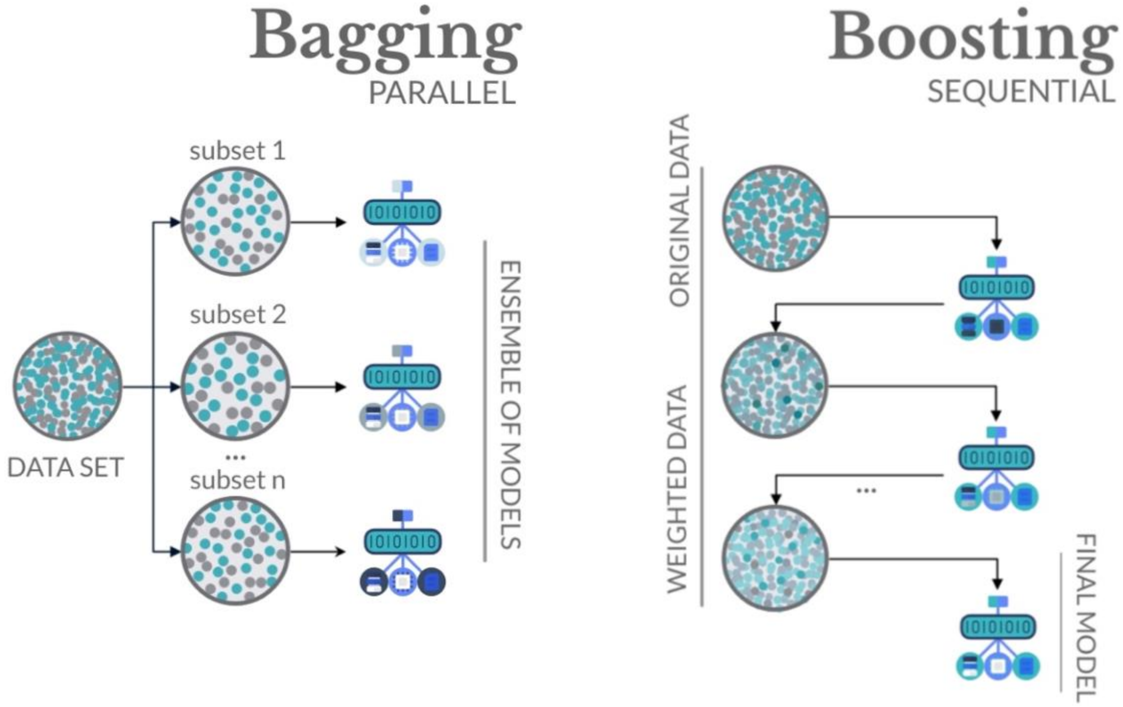


Figure 4-7 Comparison between bagging and boosting in ensemble learning algorithms.

4.3.8.4.3 RANDOM SUBSPACE

Another ensemble learning strategy is Random Subspace. Random Subspace algorithms are generally used to improve the accuracy of DFA or KNN classifiers. Like bagging, random subspace algorithms apply bootstrapping. However, as shown in Figure 4-8, random subspace algorithms apply bootstrapping to the feature space (predictor variables) of the sample whilst bagging applies bootstrapping to the sample space (Ren, Zhang & Suganthan, 2016).

fitcensemble (MathWorks, 2016) was used to train ensemble classifiers for each predictor variable or combination of predictors with all hyperparameters optimised. During sample size analysis, the template algorithm was specified as a bagged tree.

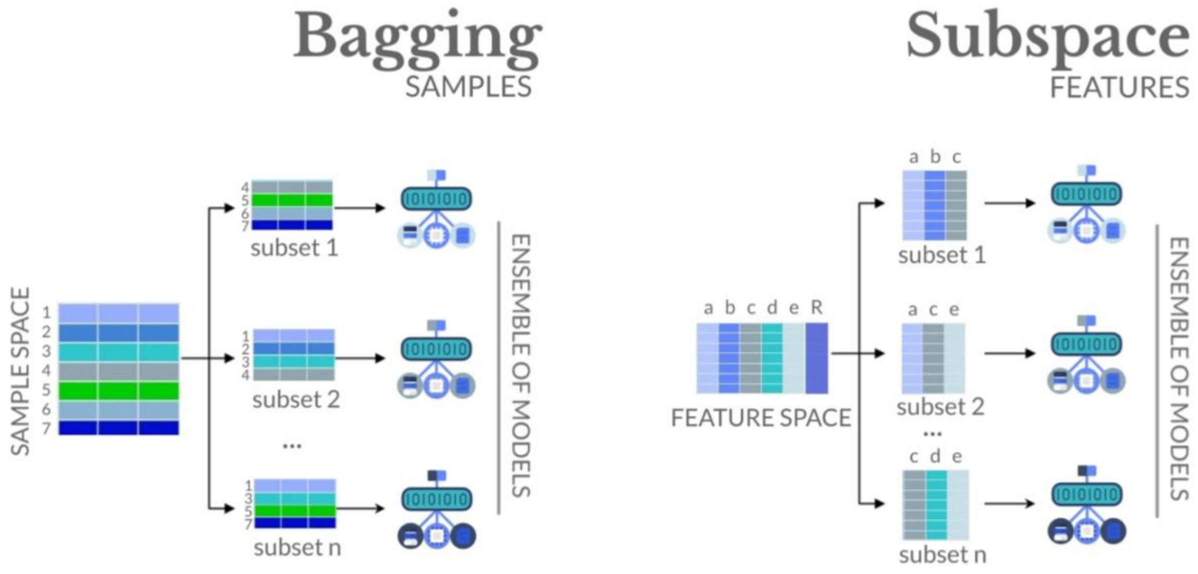


Figure 4-8 Comparison between bagging and random subspace ensemble methodologies.

4.3.8.5 DISCRIMINANT ANALYSIS

Discriminant function analysis (DFA) is commonly applied to sex classification problems in the field of biological anthropology (Steyn & Işcan, 1999; Bidmos & Dayal, 2004; Bidmos, Steinberg & Kuykendall, 2005; Dayal, Spocter & Bidmos, 2008; Soni, Dhall & Chhabra, 2010; Spradley & Jantz, 2011; Krüger, L'Abbé & Stull, 2017). The use of machine learning algorithms to perform DFA allows for quicker model training and more accurate optimization of hyperparameters than is possible when applying calculations manually (Büyüköztürk & Çokluk-Bökeoğlu, 2008).

Using LDA as an example in Figure 4-9. The premise of DFA (linear or otherwise) is to maximise the distance between class means whilst minimising the variation or scatter within the classes to reduce the number of dimensions and find the best linear (or logistic or quadratic) fit line (Büyüköztürk & Çokluk-Bökeoğlu, 2008).

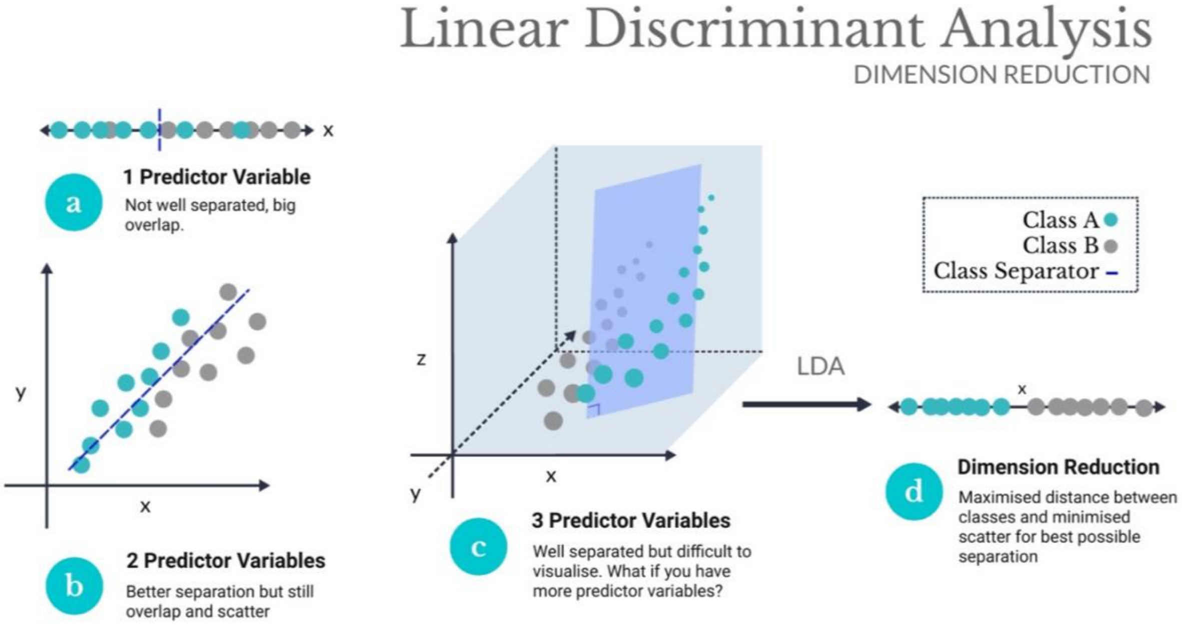

Linear Discriminant Analysis
DIMENSION REDUCTION
x
a
1 Predictor Variable
Not well separated, big overlap.
y
x
b
2 Predictor Variables
Better separation but still overlap and scatter
z
y
x
c
3 Predictor Variables
Well separated but difficult to visualise. What if you have more predictor variables?
LDA
x
Class A ●
Class B ●
Class Separator –
d
Dimension Reduction
Maximised distance between classes and minimised scatter for best possible separation

Figure 4-9 A visual explanation of the premise behind linear discriminant analysis.

Discriminant function analysis (DFA) is used to construct a discriminant function which can be used to predict the class of a new observation. DFA is based on the model that for each class (Y), data (X) follow a multivariate normal distribution. This means that DFA assumes that the data follow a Gaussian mixture distribution (MathWorks, 2007).

When making a prediction, classification is performed in a manner which seeks to minimise the expected classification cost:

$$\hat{y} = \underset{y=1,\dots,K}{\arg\ \min} \sum_{k=1}^{K} \hat{P}(k|x)C(y|k),$$

$\hat{y}$ is the predicted class of observation x. $\hat{P}(k|x)$ is the posterior probability of class k for observation x and $C(y|k)$ is the cost of incorrectly classifying an observation as y when its true class is k for K - the number of classes (in our case $K = 2$)(MathWorks, 2014d).

The *fitcdiscr* function outputs coefficient matrices which contain coefficients of the linear or quadratic boundaries between classes (MathWorks, 2014e). From these coefficient matrices, the equation of the boundary between the 2 classes is:

$$Const + Linear * x + \ x' * Quadratic * x = 0.$$

This can be used to derive a quadratic equation which can be used as a predictor for a single instance in the form:

$$y = ax^2 + bx + c,$$

where a is the '$Quadratic$' coefficient, b is the $'Linear'$ coefficient, c is $'Const'$ and $y = 0$ acts as the boundary between classes. Thus, if y > 0 for any value of x, the observation is classified as male.

For a linear function, a is simply equal to zero.

The multivariate discriminant function for variables 1 to n takes the form:

$$y = a_1x_1^2 + \ b_1x_1 + \cdots a_nx_n^2 + \ b_nx_n + e,$$

with e being the sum of all constants.

The *fitcdiscr* function (MathWorks, 2014e) was used to train models for each predictor variable or set of predictor variables. Hyperparameters were optimised.

Linear and quadratic classification functions were derived from the model coefficients for each model. These can be used to predict sex without using software.

STEP 3: ALGORITHM VALIDATION

Evaluation procedures including resubstitution, *k*-fold and holdout cross validation were implemented to aid in ensuring that study protocols were replicable, reproducible, and reliable. Training and testing subsets of the data were used to predict model accuracy and models were investigated to determine where the algorithm failed to accurately predict sex and which predictors had the biggest contribution to model success. Finally, a *post hoc* sample size analysis was performed to determine whether the dataset was sufficient in size to reliably predict sex.

4.3.9 MODEL EVALUATION PROCEDURE

The misclassification error calculated using algorithm training data is not a good estimate of how well a model will perform on new data given that the model was trained to best fit the training data. Cross-validation (CV) is a performance evaluation technique used to test the efficacy of ML algorithms when applied to new datasets that they were not trained on.

When training a supervised classification model, overfitting by creating an overcomplex model or, underfitting by training an overly simplistic model are risks that need to be mitigated. Over and underfitting can be identified via cross-validation, the splitting of a dataset into subsets which are then used to calculate how well the model performs on new data.

CV can be performed using several techniques. In the case of this study, 3 techniques were used in parallel to compute the accuracy of models. First, the dataset was partitioned into two subsets for training and testing.

The training subset was used for *resubstitution,* this involved using the training data to evaluate model performance. This is expected to produce overly optimistic results but is a good measure for over- or under-fitting when compared to other cross-validation results.

The testing subset was a *holdout* (Ho) sample of 20% of the dataset. This data was not involved at all in training but rather retained for algorithm validation after training was complete. The sample size of this study is relatively small compared to standard machine learning datasets, so the Ho accuracy is expected to be highly variable given that the sample is small and the population highly heterogenous. Used in isolation, holdout validation is highly susceptible to the dataset, especially with smaller samples (MathWorks, 2021d). Hence, it was used in combination with resubstitution and *k*-fold cross-validation.

Finally, *k*-fold cross-validation was used to evaluate model performance. *K-fold cross-validation* partitions the training data into *k* randomly selected, similarly sized subsets. For each subset or 'fold', k-1 subsets are pooled to for training and the remaining subset is used to validate the model. This process is repeated *k* times so that each subset is used to validate the model exactly once.

K-fold cross validation was performed using 10 folds for each model, meaning that the data were partitioned into 10 subsets and each set was reserved as a validation set whilst the 9 other sets were used to train a model (MathWorks, 2012). The accuracies of these 10 models were then averaged and presented as an accuracy estimate. K-fold CV accuracy is the best true representation of model accuracy as it averages the result across multiple folds and reduces the variance between accuracy values. Therefore k-fold cross validation accuracies are used to compare performance of models rather than training accuracies or Ho accuracies which are both far more susceptible to small differences in the dataset. The larger variability in estimated model accuracy at smaller sample sizes will be explained in a forthcoming section.

Cohen's kappa (Layden, 2019) was also calculated for each classification model and cross-validation type. Kappa (κ) calculates percentage agreement, considering the possibility of agreement occurring by chance. A κ statistic between 0.61 and 0.80 corresponds with substantial agreement and any κ statistic exceeding 0.81 is almost perfect agreement (McHugh, 2012).

Overfitting occurs when a model includes uses an overcomplicated approach with more predictors or coefficients than needed to solve a machine learning task (Hawkins, 2004). A good machine learning model is ideally as simplistic as possible whilst remaining accurate and being able to produce meaningful predictions. More complex models are more prone to overfitting and as such, avoiding training of models with more coefficients and predictors than necessary to describe structure of the data is a strategy to mitigate overfitting. Another measure to avoid overfitting is ensuring that there is sufficient training data. Sample sizes which are too small for a particular algorithm or modelling method will yield results which may not translate well to new data. Once a model has been trained, overfitting is easily identified using cross-validation (Roelofs et al., 2019). If the prediction accuracy of a model decreases significantly between the training accuracy and *k*-fold CV accuracy, then it is likely that the model has been overfit.

Overfitting was identified by comparing the training accuracy of a model to the *k*-fold CV accuracy. If the training accuracy was markedly higher, then the model was considered to be overfit. If the training accuracy was 100% but the *k*-fold CV accuracy and testing were lower then, the model was conclusively overfit.

4.3.10 *POST HOC* SAMPLE SIZE AND PREDICTOR IMPORTANCE ANALYSIS

There is little consensus on what constitutes a large sample size compared to a small sample size. In fact, this definition varies depending on the type of algorithm being referred to as well as the intended application. It is easy to label a sample size of 15 (Golland et al., 2000), 30 (Patel et al., 2013) or 58 (Henderson & Nikita, 2016) as small and 1 000 or even 500 000 (Vabalas et al., 2019) as large. But sample size in relation to the predictive outcomes is far more important than size as a fixed number. The goal of machine learning is to develop models which can generalise to make predictions on new data (Baum & Haussler, 1989). So how do we determine the optimal sample size to fulfil this objective?

A review of 167 articles pertaining to machine learning in the field of medical imaging found only 4 articles which performed any form of *pre hoc* sample size determination and a paltry 18 attempted *post hoc* sample size determination. *Pre hoc* methods tended to yield prohibitively large sample size estimates, often exceeding 4 000 observations per class, or made many (often unrealistic) assumptions about the underlying structure of the data and, failed to consider algorithm type (Balki et al., 2019). It is certainly wrong to assume that the required dataset size is the same for all types of machine learning algorithm but, as yet there is no *pre hoc* formulaic method for determining exact sample sizes for a given model.

The most common *post hoc* method for determining sample size is plotting a learning curve. This involves taking increasingly large subsets of a dataset, training a model, and calculating the error before plotting this on a curve (Vabalas et al., 2019). This can then provide the sample size needed for a particular error rate and has the added benefit of being helpful to future researchers. In general, the variability of performance estimates tends to be greater with smaller sample sizes and more consistent as sample sizes increase (Vabalas et al., 2019).

4.3.10.1 OPTIMAL NUMBER OF PREDICTORS

The most relevant, least redundant predictors were identified using the MRMR algorithm as well as other indicators of predictor success in the results. These predictors were used in order of highest importance to train models with an increasing number of predictors. Each model was trained 1000 at sample sizes varying from 10 to 400 and mean model errors at each sample size were plotted on a curve.

4.2.10.2 OPTIMAL SAMPLE SIZE

The optimal number of predictors was determined using the curve described above. Consequently, 3-predictor models were trained 1000 times each at increasingly large sample sizes for each algorithm type. Mean accuracies as well as upper and lower boundaries for a 95% confidence interval (CI) were plotted for each algorithm type. These were used to evaluate optimal sample size of a 3-predictor model for each algorithm type

4.3.11 PREDICTOR IMPORTANCE

Embedded type feature importance was calculated for selected multivariate models. When assessing discriminant models, Delta values for predictors were used to rank predictor importance and contributed to the model’s success. For decision trees and ensemble trees, the ‘predictorImportance’ command was used to extract and plot the contribution of each predictor to the model.

CHAPTER 5: RESULTS

5.1 BASELINE ANALYSIS

5.1.1 OBSERVER AGREEMENT RESULTS

The results of observer agreement analysis using Lin's CCC are presented in Table 5-1. All the left- and right-sided measurements for the 12 mensurations were in agreement for both inter- and intra-observer analysis. The intra-observer scores were higher than those for inter-observer. None of the scores were lower than 0.950. The lowest score was for RTAD inter-observer agreement and was 0.959. Figure 5-1 further explores the results for RTAD. A single outlier, for which coordinates are given is present. Table 5-2

Table 5-1 Inter-Observer and Intra-Observer CCC scores for all left- and right-sided mensurations.

Left	**CCC**		Right	**CCC**	
Mensuration	Inter-	Intra-	Mensuration	Inter-	Intra-
LAD	0.978	0.997	**RAD**	0.981	0.997
LTAD	0.966	0.997	**RTAD**	0.959	0.997
LFHD	0.995	0.998	**RFHD**	0.996	0.999
LTHD	0.997	0.999	**RTHD**	0.994	0.999
LMCL	0.982	0.996	**RMCL**	0.967	0.996
LBB	0.993	0.999	**RBB**	0.996	0.998
LLCL	0.992	0.997	**RLCL**	0.989	0.997
LTPB	0.981	0.995	**RTBP**	0.978	0.993
LGL	0.987	0.996	**RGL**	0.976	0.994
LGB	0.985	0.994	**RGB**	0.968	0.990
LEB	0.968	0.996	**REB**	0.994	0.999
LHHD	**0.954**	0.996	**RHHD**	0.985	0.995

CCC – Lin's concordance correlation coefficient.
Notable data highlighted with **bold** blue text.

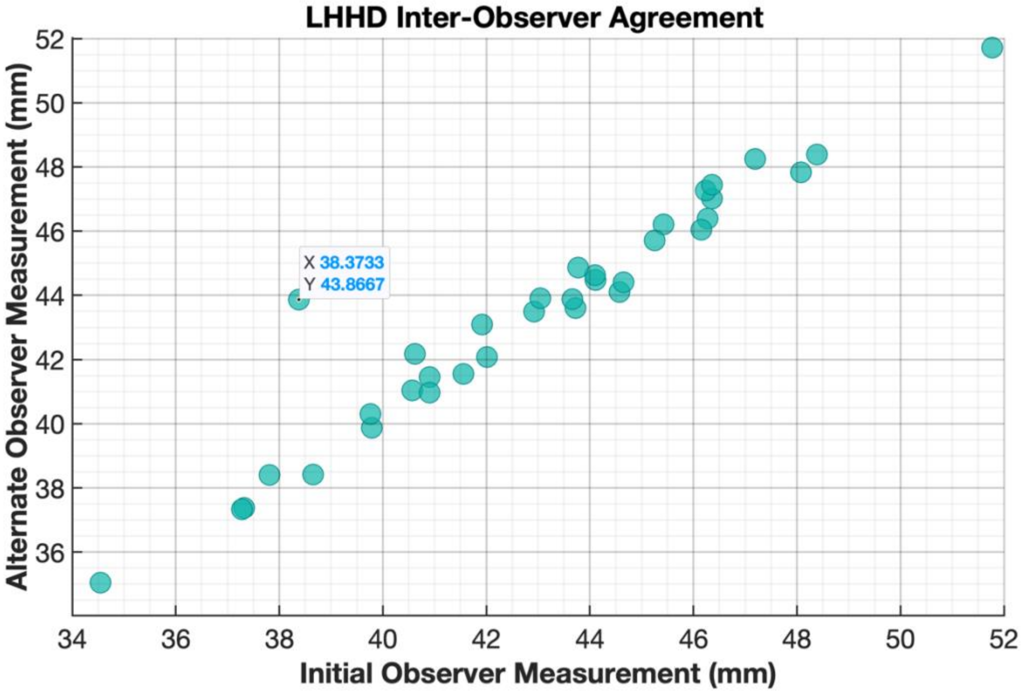


Figure 5-1 Scatter plot of measurements taken by initial observer compared to those taken by a second observer for LHHD. An outlying datapoint [38.37, 43.87] Is highlighted.

Table 5-2 LHHD Inter-Observer measurement averages for outlying data point [38.87, 43.37] in Figure 5-1.

	Initial Observer	Inter-observer
Parameter	mm	mm
Overall Mean	41.12	41.30
Left Mean	38.37	43.87
Right Mean	43.86	38.72

5.1.2 OUTLIERS RESULTS

Boxplots for each mensuration denote outliers as a red ‘x’ (Figure 5-2&3). Some predictors, like MCL and BB, had very few outlying datapoints whilst others, like TPB, GL and EB, had many outlying datapoints. Outliers were often close to the boxplot tail. In these cases, the outliers most likely did not have a large influencing effect on dataset parameters or predictive models. Conversely, the extreme female outlier in the HHD boxplot, for example, may have an influencing effect on dataset parameters, like means, if the sample size is small.

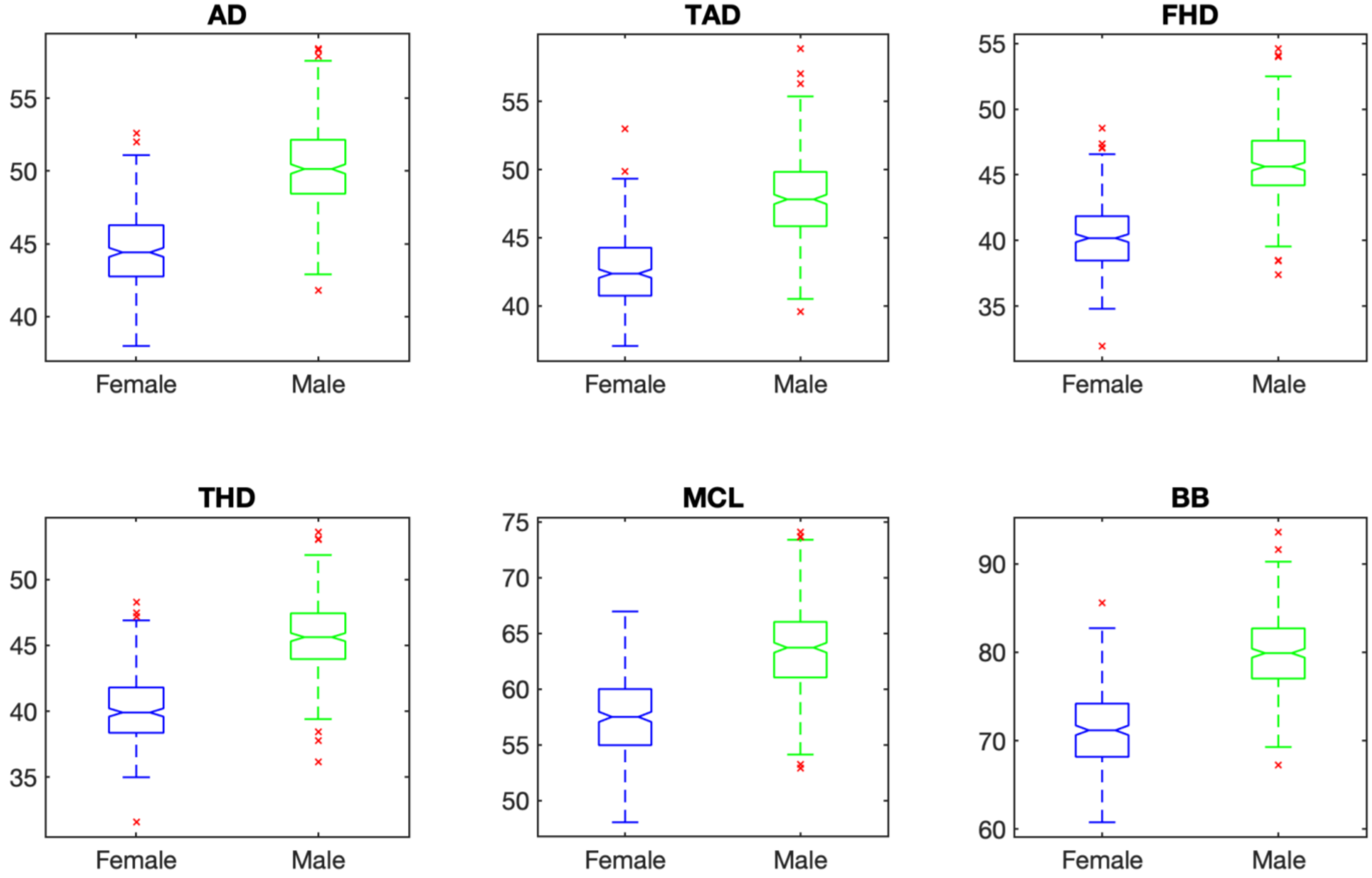


Figure 5-2 Boxplots for AD, TAD, FHD, THD, MCL and BB highlighting male and female outliers as red 'x' markers.

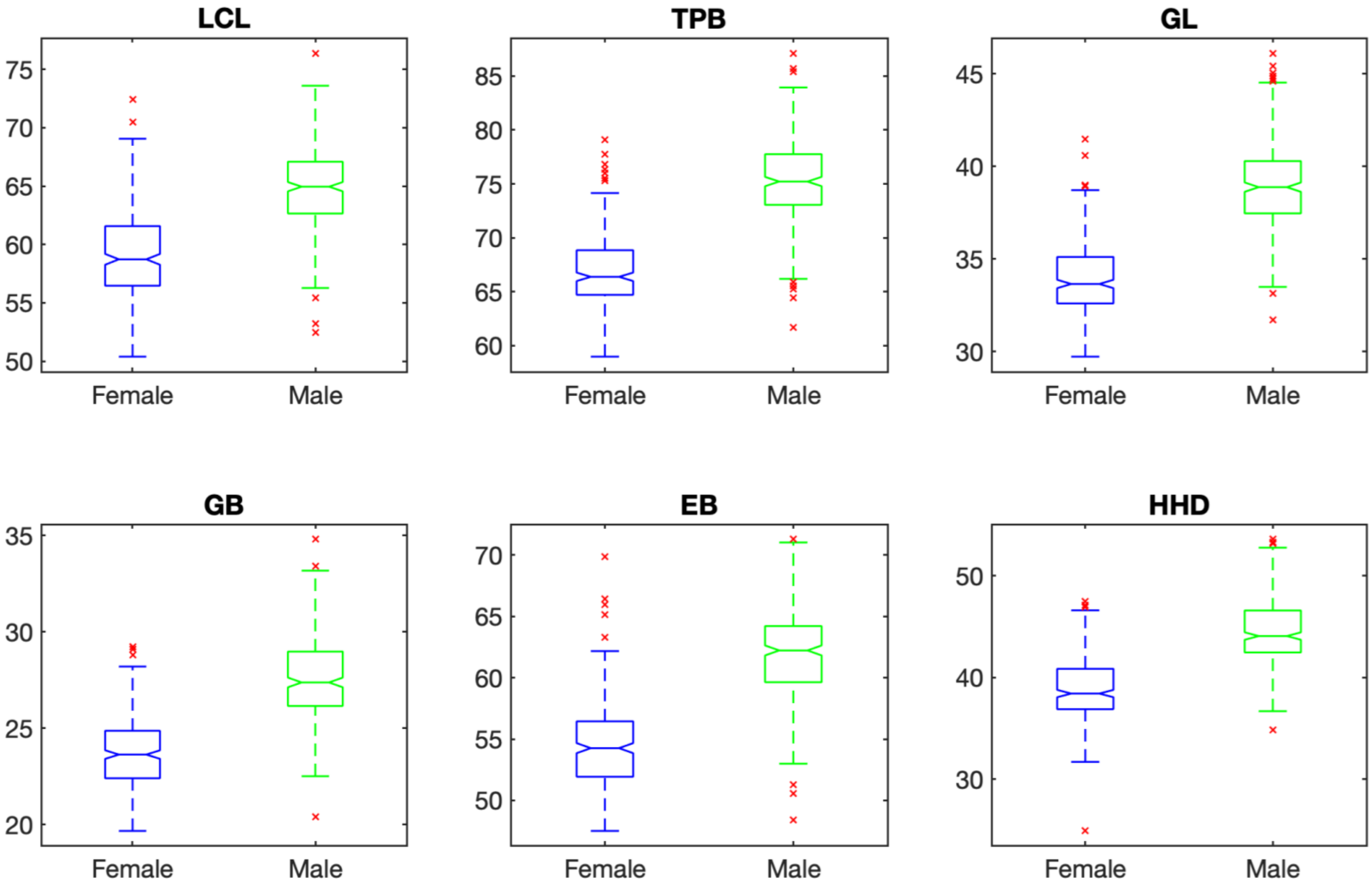


Figure 5-3 Boxplots for LCL. TPB, GL, GB, EB, and HHD highlighting male and female outliers as red 'x' markers.

5.1.3 NORMALITY TESTING RESULTS

Table 5-3 contains the results of statistical testing for normality. When applying the *kstest*, all mensurations had a p-value <0.05. These results suggested that none of the mensurations were normally distributed. In comparison, results from the *jbtest* indicated that LMCL, RMCL and LHHD were normally distributed. These mensurations had p-values >0.05 and thus the alternate hypothesis that the data are not normally distributed was rejected. Given these results, all mensurations; except LMCL, RMCL and LHHD, were conclusively not normally distributed.

A further visual assessment was performed on the 3 uncertain parameters. LMCL (Figure 5-4) had a cluster of tall bars around the 55mm mark. RMCL in Figure 5-5 appeared to be normal, however in Figure 5-6 where more bins were added, the many extreme values in the data were more visible. RHHD (Figure 5-7) did not appear to be typically normally distributed.

Table 5-3 Results of statistical testing for normality using the kstest and jbtest.

	kstest	***jbtest***		***kstest***	***jbtest***
Mensuration	p	p	Mensuration	p	p
LAD	<0.001	0.005	**RAD**	<0.001	0.006
LTAD	<0.001	0.003	**RTAD**	<0.001	0.007
LFHD	<0.001	0.018	**RFHD**	<0.001	0.010
LTHD	<0.001	0.014	**RTHD**	<0.001	0.008
LMCL	<0.001	**0.143**	**RMCL**	<0.001	**0.083**
LBB	<0.001	0.008	**RBB**	<0.001	0.010
LLCL	<0.001	**0.052**	**RLCL**	<0.001	0.026
LTPB	<0.001	0.005	**RTBP**	<0.001	0.007
LGL	<0.001	0.003	**RGL**	<0.001	0.003
LGB	<0.001	0.017	**RGB**	<0.001	0.004
LEB	<0.001	0.002	**REB**	<0.001	0.003
LHHD	<0.001	0.013	**RHHD**	<0.001	**0.339**

jbtest – Jarque Bera Test for Normality, ***kstest*** – Kolmogorov Smirnov Test for Normality.
*Values with p>0.05 highlighted with **bold** blue text.

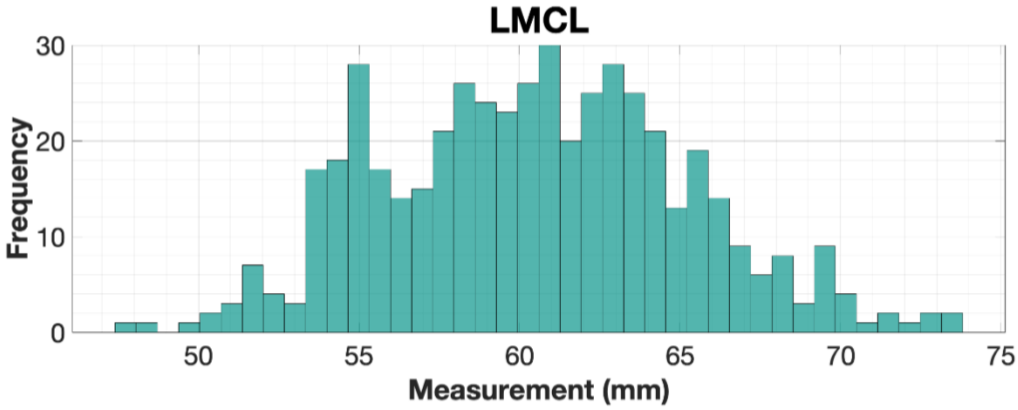


Figure 5-4 Histogram of measured values for LMCL.

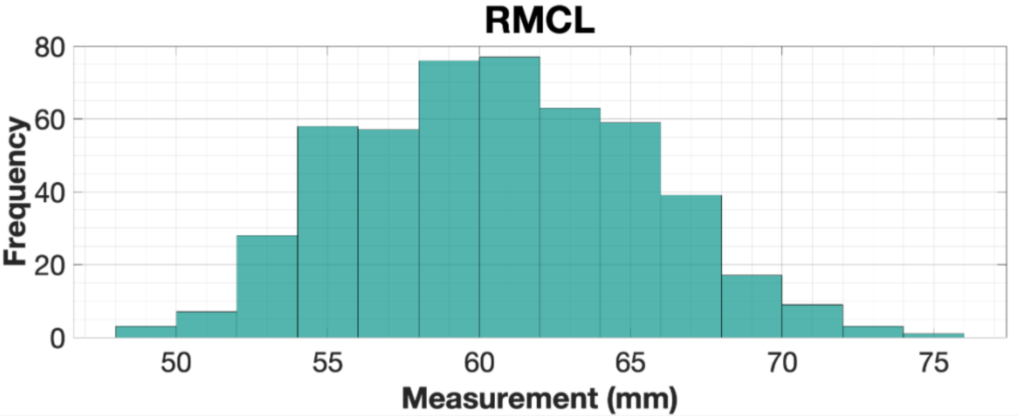


Figure 5-5 Histogram of measured values for RMCL with number of bins set to default.

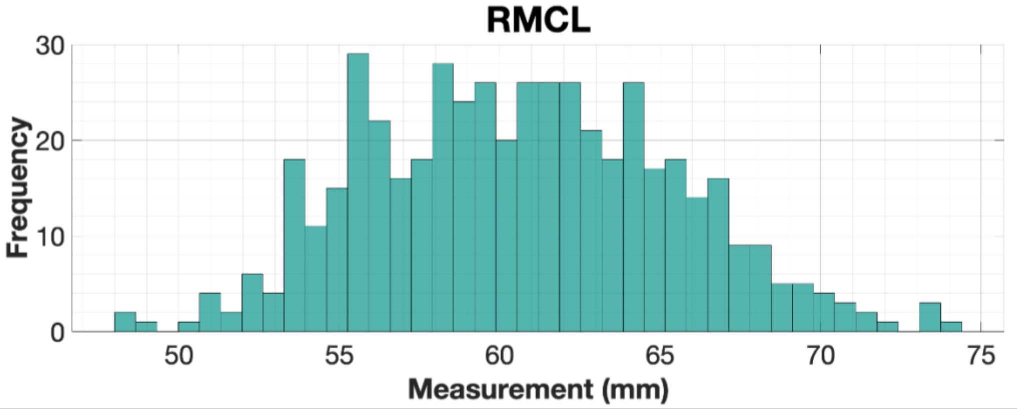


Figure 5-6 Histogram of measured values for RMCL with number of bins set to 40.

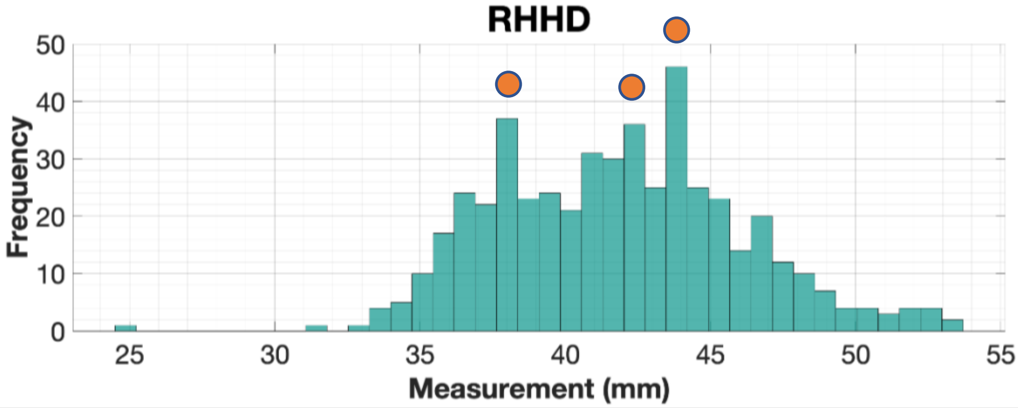


Figure 5-7 Histogram of measured values for RHHD with orange dots highlighting extreme peaks.

5.2 RESULTS OF DEMOGRAPHIC ANALYSIS

5.2.1 BILATERAL ASYMMETRY

The Wilcoxon Signed Rank test found no significant differences between left- and right-sided measurements for all mensurations except GB (Table 5-4).

GB was the smallest measurement and had a difference between left- and right-sided median values of 0.13mm. This was smaller than the standard error associated with this mensuration. 74.75% of all left- and right- paired measurements for GB fell within 1mm of each other (Table 5-5). Visual inspection of Figure 5-8 shows that there is very little difference between left- and right-sided measurements for GB.

The difference between left- and right-side medians, as presented in Table 5-4, ranged from 0.02mm to 0.45mm. In most cases, this was less than standard error. These results were mirrored in Table 5-5 where the majority of paired values were within 1mm of each other.

GL was the only mensuration for which the one side is significantly larger than the other however, the median difference, once SE was considered, was only 0.29mm.

TAD, GL, and EB had the largest proportion of pairs where one side (in all cases, the right side) was larger than the other. Violin plots for these mensurations (Figure 5-8) showed that there was very little obvious difference between left- and right-sided probability distributions.

Table 5-4 Comparison between left- and right-sided median measurements and results of signrank test.

	Median ± SE			***signrank***
Mensuration	L	R	Difference	p
AD	47.22 ± 0.19	47.31 ± 0.19	-0.09	0.026
TAD	44.87 ± 0.19	45.16 ± 0.19	-0.29	<0.001
FHD	42.91 ± 0.20	42.93 ± 0.19	-0.02	<0.001
THD	42.66 ± 019	42.88 ± 0.19	-0.22	<0.001
MCL	60.47 ± 0.24	60.59 ± 0.24	-0.12	<0.001
BB	75.33 ± 0.31	75.76 ± 0.31	-0.43	<0.001
LCL	62.14 ± 0.24	61.84 ± 0.24	0.30	0.023
TPB	70.95 ± 0.31	71.00 ± 0.31	0.05	<0.001
GL	35.97 ± 0.16	36.42 ± 0.16	-0.45	<0.001
GB	25.44 ± 0.14	25.31 ± 0.14	0.13	**0.696**
EB	58.00 ± 0.26	58.06 ± 0.27	-0.06	<0.001
HHD	41.66 ± 0.21	41.82 ± 0.22	-0.16	<0.001

SE – standard error, ***signrank*** – Wilcoxon Signed Rank test for agreement between paired values.
*Value with p>0.05 highlighted in **bold** blue text**.** Largest differences in median highlighted in cyan.

Table 5-5 Percentage of pairs with bilateral differences between paired measurement values exceeding 1mm.

	Percentage of pairs with:		
Measurement	R > L	R < L	R = L
AD	10.58%	16.32%	73.10%
TAD	24.54%	13.30%	**62.16%**
FHD	7.57%	2.61%	89.81%
THD	6.27%	3.13%	90.60%
MCL	21.47%	13.61%	64.82%
BB	22.05%	6.56%	71.39%
LCL	11.00%	15.71%	73.30%
TPB	20.06%	10.49%	69.44%
GL	30.10%	6.97%	**62.94%**
GB	13.00%	12.25%	74.75%
EB	28.01%	12.04%	**59.95%**
HHD	17.85%	7.09%	75.07%

Smallest percentage of R (right) = L (left) pairs highlighted in **bold** blue text.

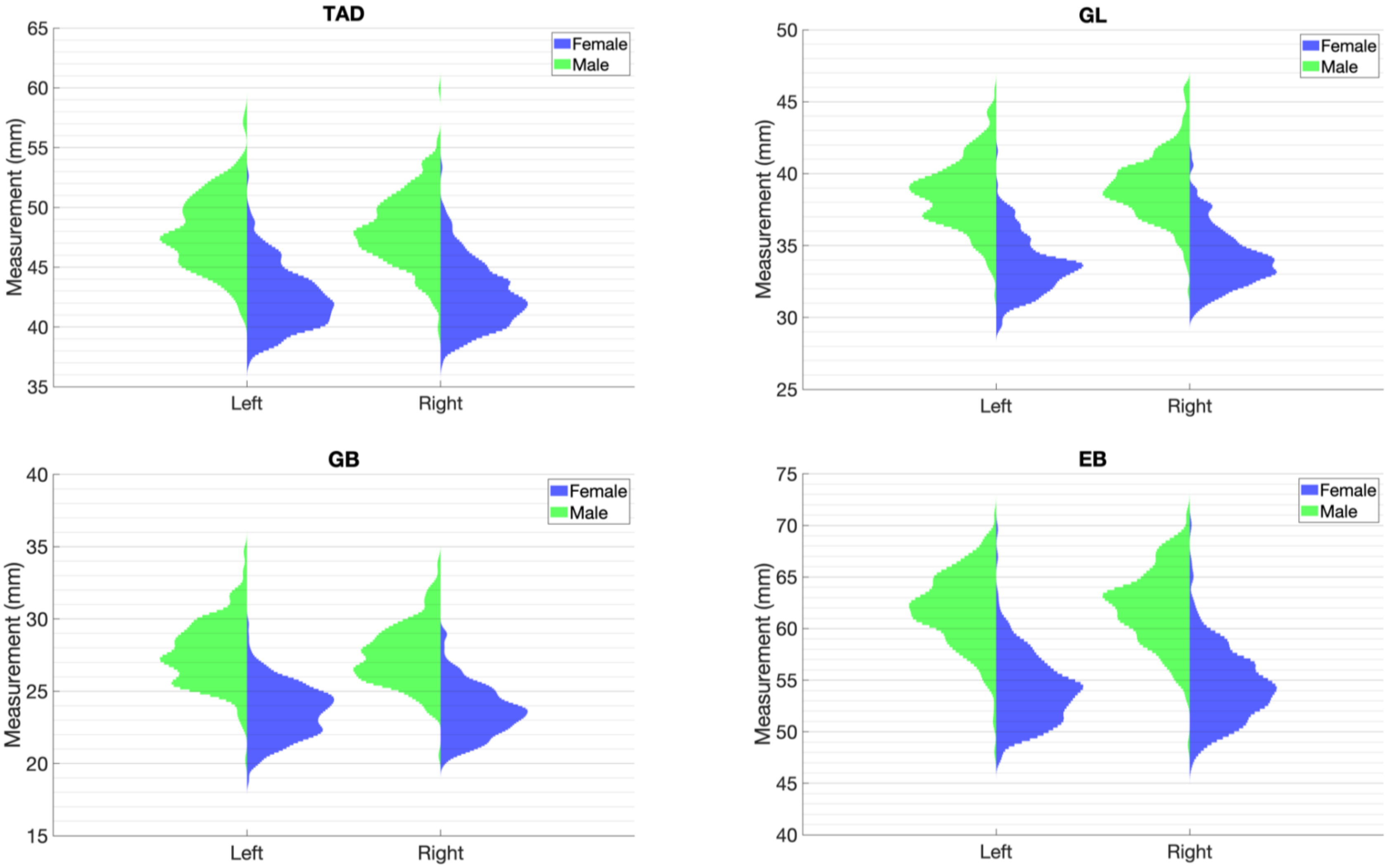


Figure 5-8 Violin plots comparing left- and right-sided TAD, GL, GB, and EB measurements.

5.2.2 SEXUAL DIMORPHISM RESULTS

Results of testing for sexual dimorphism are presented in Table 5-6. The p-value for all mensurations was much smaller than 0.05 and thus the null hypothesis was rejected for all mensurations. Therefore, all mensurations were sexually dimorphic. MCL and LCL had the biggest p-values and the smallest percentage difference between male and female values. GL and GB had the smallest p-values and the largest percentage difference between the sexes. As a result of these statistics, MCL and LCL were likely the least dimorphic measurements and, GL and GB were likely the most dimorphic measurements.

Table 5-6 Results of ranksum test for sexual dimorphism and male and female descriptive statistics.

	Male		**Female**			***ranksum***
Measurement	Median	Variance	Median	Variance	% Difference	p
AD	50.14	8.46	44.41	6.23	11.46	1.45×10^{-75}
TAD	47.82	8.91	42.37	6.59	11.40	2.47×10^{-71}
FHD	45.63	8.18	40.18	6.20	11.94	2.66×10^{-74}
THD	45.63	7.99	39.91	6.02	11.54	7.09×10^{-75}
MCL	63.71	14.21	57.51	12.64	9.73	**1.49×10^{-59}**
BB	79.91	19.14	71.17	16.94	10.94	1.02×10^{-72}
LCL	64.95	12.82	58.73	14.50	9.58	**8.83×10^{-54}**
TPB	75.21	14.33	66.37	11.72	11.75	7.74×10^{-74}
GL	38.87	5.08	33.64	3.78	13.46	6.92×10^{-83}
GB	27.37	4.35	23.63	2.99	13.66	1.95×10^{-76}
EB	62.22	12.92	54.26	11.40	12.79	5.43×10^{-77}
HHD	44.08	9.80	38.43	8.44	12.82	1.22×10^{-70}

ranksum – Wilcoxon rank sum test.
Largest p-values highlighted in **bold** blue text and smallest p-values are highlighted in cyan

The 4 most extreme mensurations, MCL, BB, GL and GB were further examined in the graphs that follow. The degree of overlap between male and female values is visibly smaller for GL (Figure 5-11) and GB(Figure 5-12) than for MCL Figure 5-9) and LCL (Figure 5-10). The frequencies for the modal bars in GL and GB are higher whilst the data are more spread out over a larger range of values with weaker clustering about the mode for MCL and LCL.

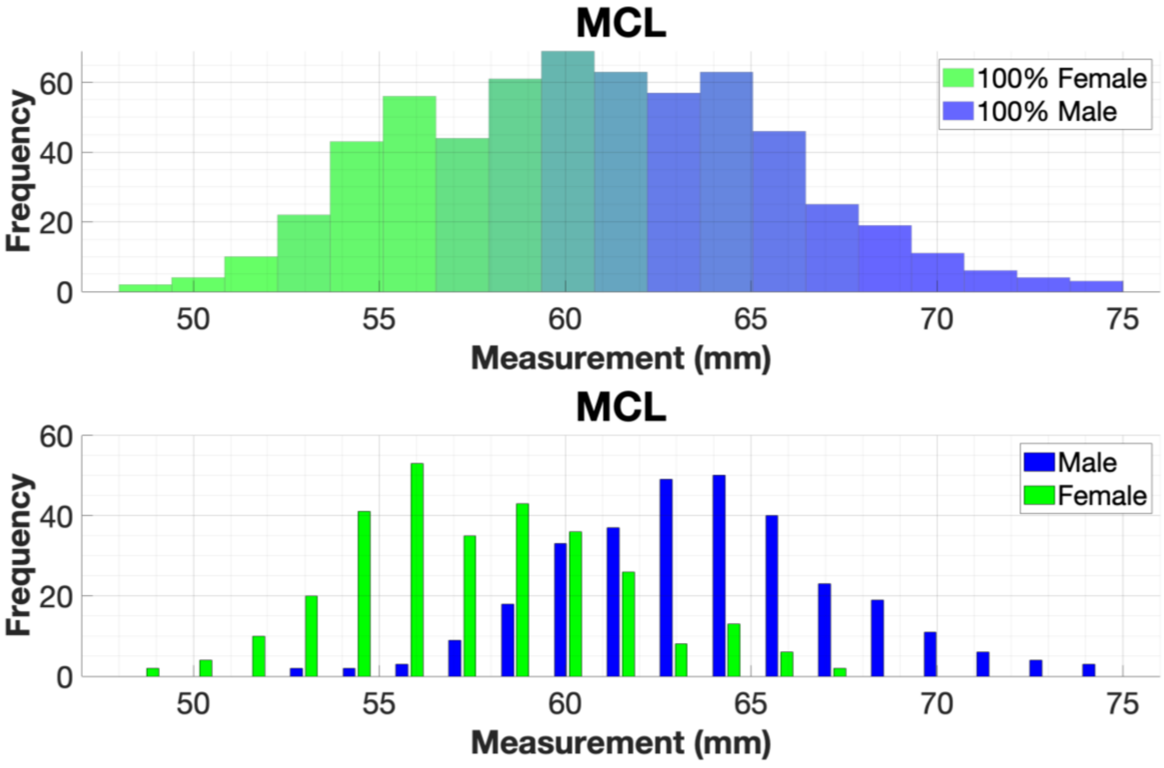


Figure 5-9 Blended histogram and bar chart showing male and female measurement values for MCL.

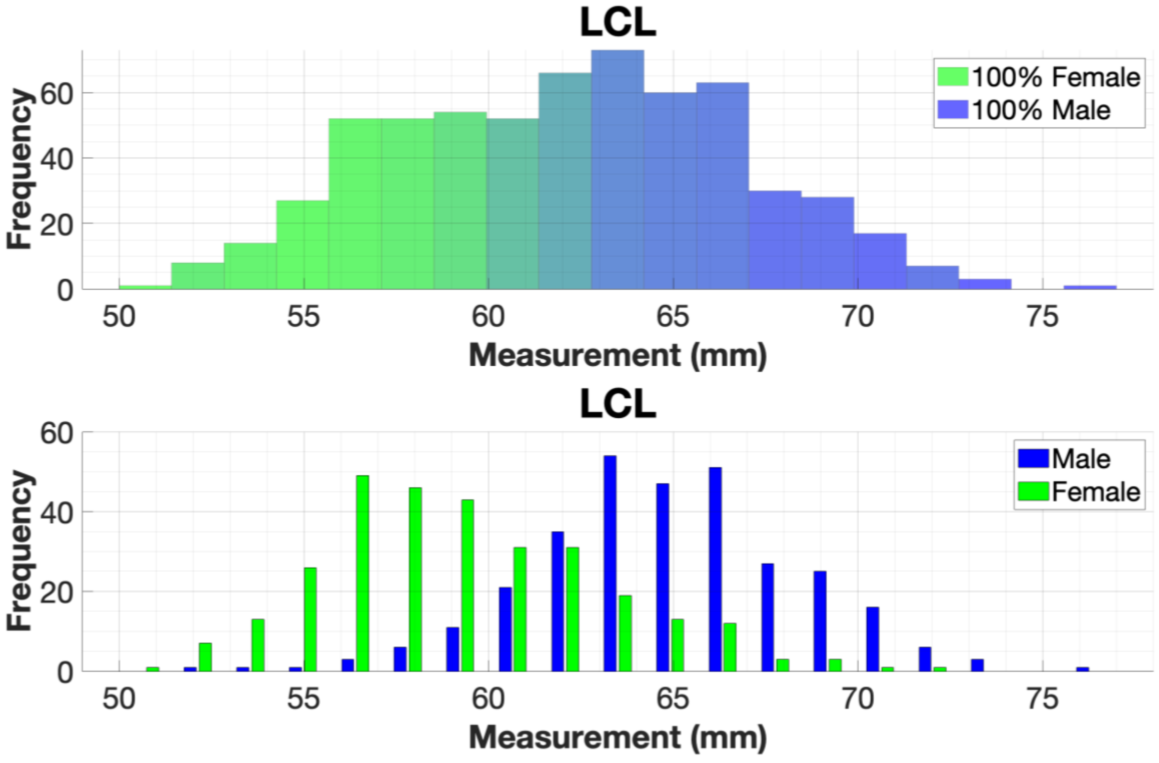


Figure 5-10 Blended histogram and bar chart showing male and female measurement values for LCL.

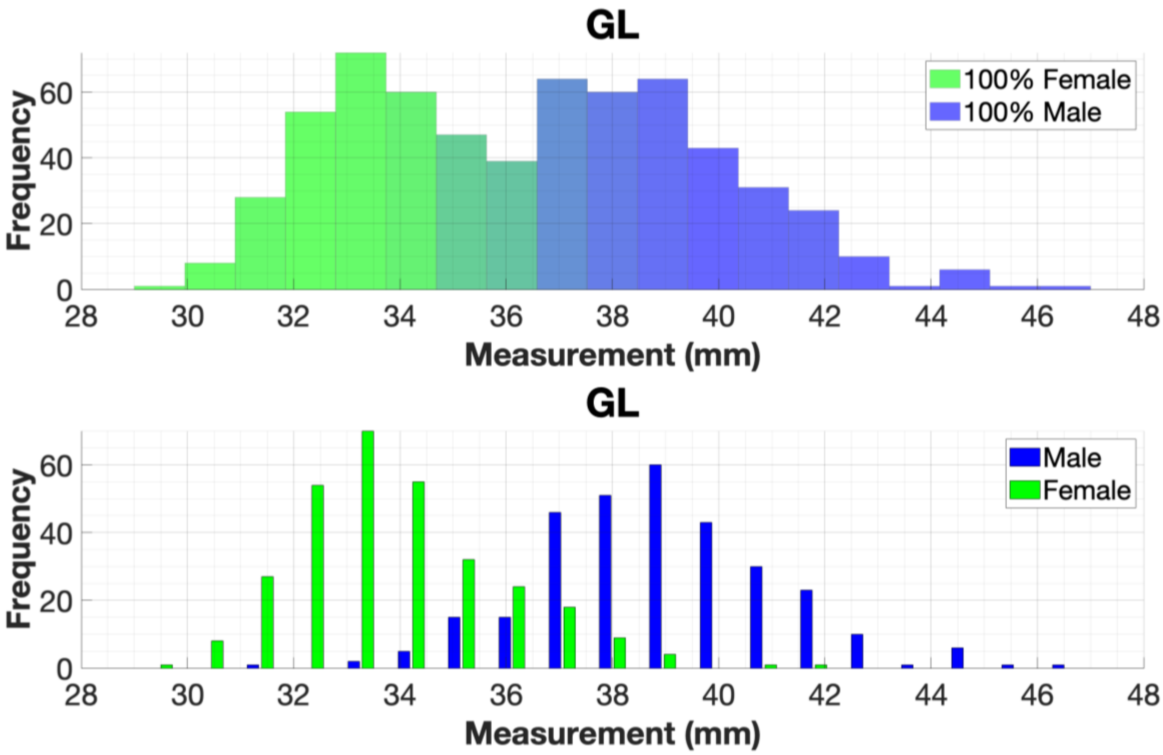


Figure 5-11 Blended histogram and bar chart showing male and female measurement values for GL.

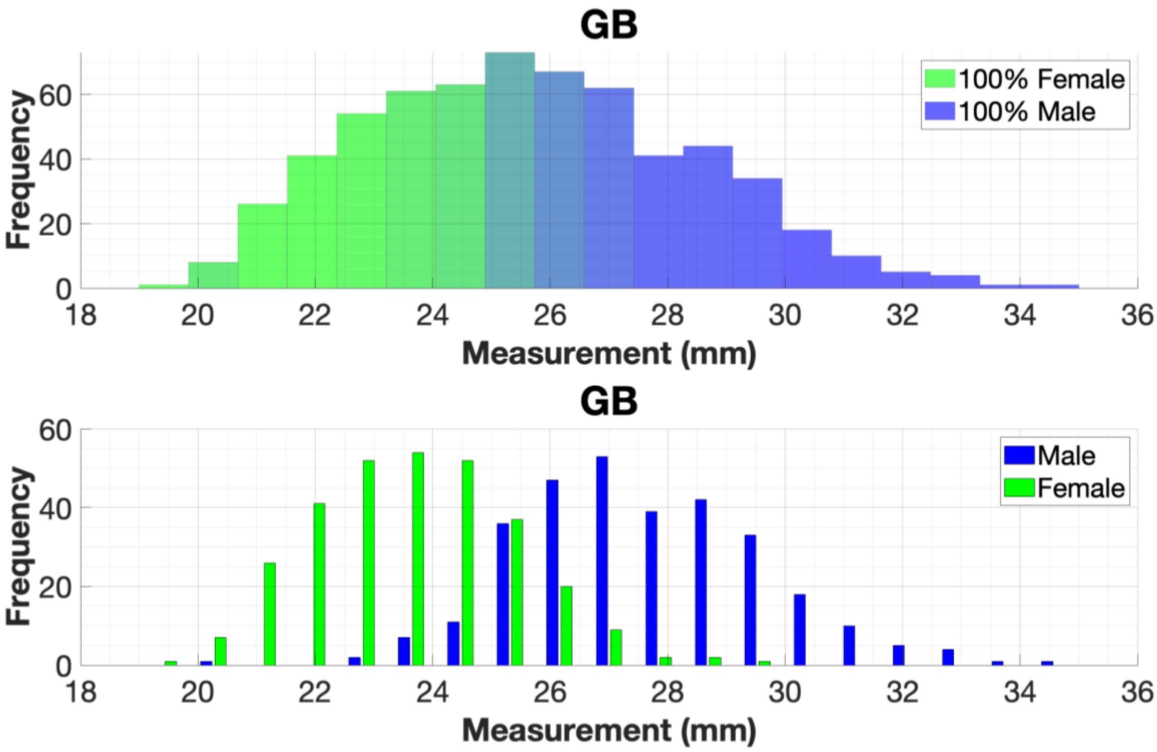


Figure 5-12 Blended histogram and bar chart showing male and female measurement values for GB.

5.2.3 AGE AND SECULAR TREND RESULTS

The study sample consisted of male and female individuals with an age-at-death between 18 and 65 years who were born between 1875 and 1992.

Figure 5-13 shows the distribution of individuals by year of birth. The earliest year-of-birth was 1875 and the most recent was 1992. The mean age-at-death for males (48) was higher than for females (42) although the age range was the same for both sexes. An effort was made during sampling to ensure that the study sample was both as modern as possible (most recent year-of-birth) and as young as possible (lowest age-at-death).

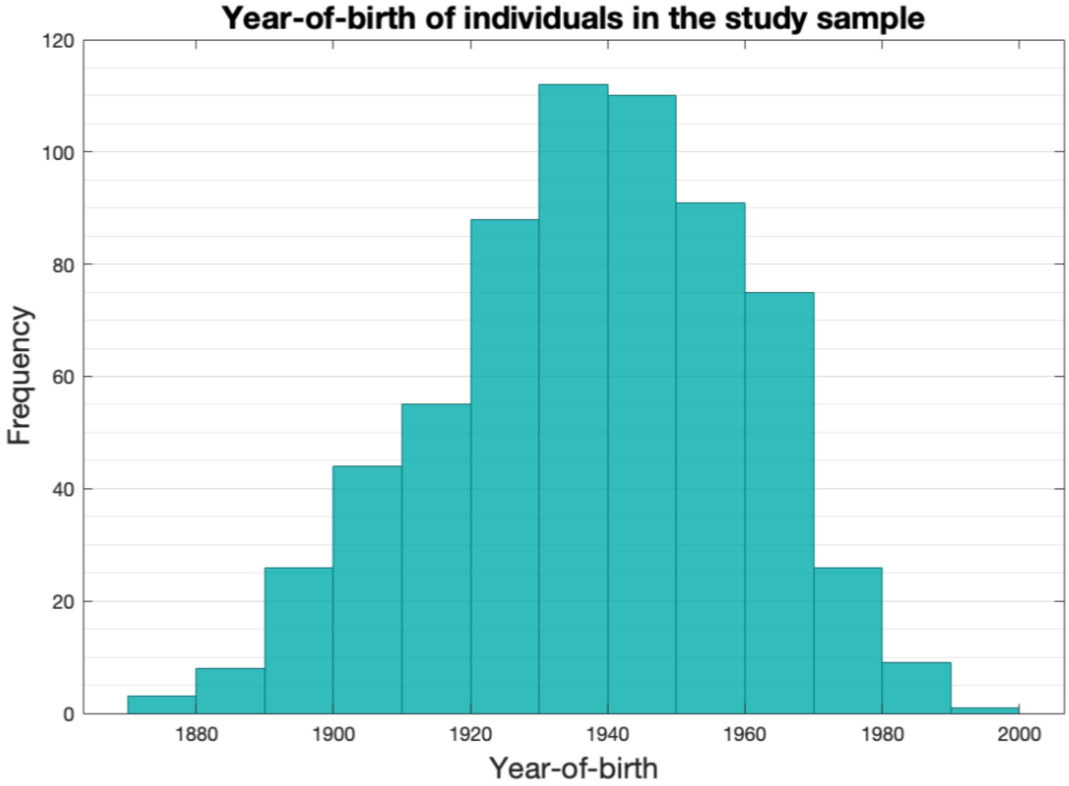


Figure 5-13 Histogram showing Year-of-birth distribution for individuals in the study sample.

Due to the nature of skeletal collections, many are skewed towards containing more individuals with a higher age-at-death. The Dart skeletal collection, as seen in Figure 5-14, had a good number of individuals in each age-group whereas the UCT skeletal repository was severely biased towards older individuals (Dayal et al., 2009; Maass & Friedling, 2019). The study sample was well-distributed when compared to the South African age-at-death distribution with most individuals in the study sample having an age-at-death between 40 and 60.

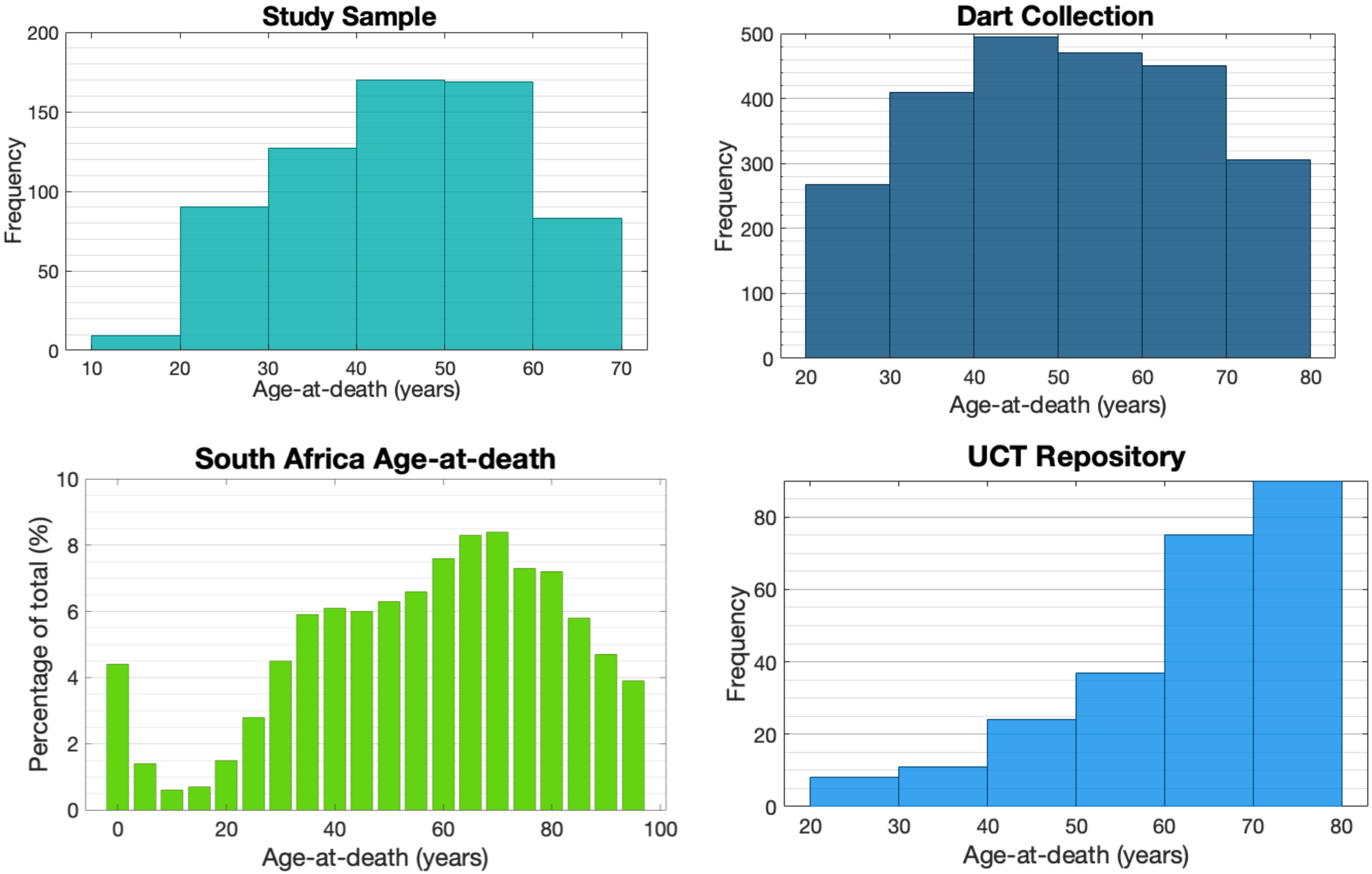


Figure 5-14 Age-at-death distributions for the study sample compared to Dart and UCT skeletal collections and the South African population estimate (2019).

Secular trend was assessed using Pearson's *r* by measuring correlation between measurement values and year-of-birth and the results are presented in Table 5-7. There was no evidence of correlation between female measurements and year-of-birth (*r*-values very small and p-values very large). Weak negative linear correlation was present between male measurements and year-of-birth for all mensurations except LCL.

Pearson's *r* was also used to measure correlation between measurement value and age-at-death. There was significant (p<0.05) but relatively weak positive linear correlation for all mensurations for both males and females (Table 5-7). The strongest correlation was for male GB (*r* = 0.40) followed by female GB (*r* = 0.36).

Table 5-7 Assessment of correlation between Year-of-birth (YOB) or Age-at-death (AAD) and measurement using Pearson's r.

	YOB				AAD	
	Male		Female		Male	Female
Mensuration	*r*	p	*r*	p	*r*	*r*
AD	-0.19	<0.01	0.02	0.79	0.25	0.25
TAD	-0.24	<0.01	0.02	0.75	0.31	0.31
FHD	-0.25	<0.01	-0.02	0.78	0.26	0.24
THD	-0.21	<0.01	0.02	0.74	0.23	0.22
MCL	-0.11	0.04	-0.02	0.69	0.18	0.18
BB	-0.19	<0.01	-0.02	0.72	0.26	0.26
LCL	-0.09	**0.10**	0.04	0.47	0.18	0.15
TPB	-0.19	<0.01	0.03	0.61	0.27	0.24
GL	-0.17	<0.01	-0.03	0.61	0.28	0.23
GB	**-0.30**	<0.01	-0.12	0.05	0.40	0.36
EB	-0.14	0.01	-0.04	0.49	0.25	0.14
HHD	-0.23	<0.01	0.01	0.87	0.24	0.28

p< 0.05 for all male and female mensurations for AAD.
YOB: Year-of-birth, AAD: Age-at- death.
Significant value highlighted in **bold** blue text.

5.3 SEX CLASSIFICATION MODELLING RESULTS

Feature importance, statistics for training and testing subsets of the data, classification model accuracies and a *post hoc* sample size analysis will be presented in this section.

5.3.1 FEATURE IMPORTANCE

According to predictor importance scores calculated using the MRMR (maximal relevance and minimal redundancy) algorithm, results suggested that when all 12 variables were combined, GL and TPB had the highest importance whilst LCL and FHD had the lowest scores. A model combining predictors with high scores would be expected to have the maximum accuracy whilst avoiding the most redundancy.

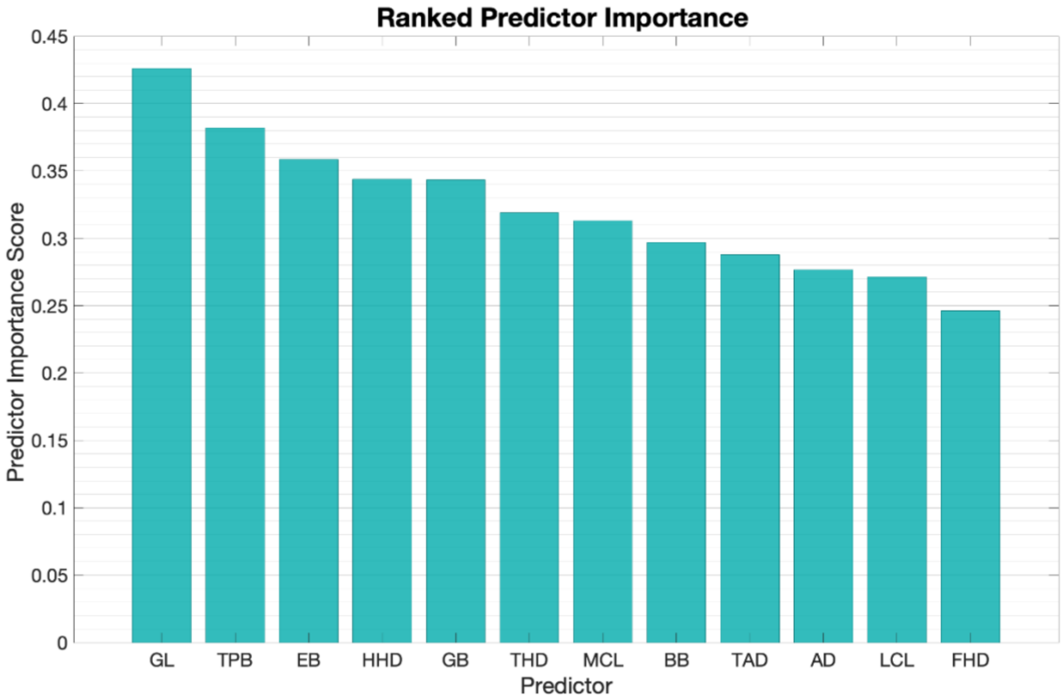


Figure 5-15 MRMR predictor importance ranking.

5.3.2 DESCRIPTIVE STATISTICS

Descriptive statistics including median, range, standard deviation (SD) and standard error (SE) are presented for each predictor (Table 5-8). Standard error was small (0.01 to 0.02) for all predictors. All male medians were larger than female medians. There was a large degree of overlap between male and female ranges for all predictors. BB had the largest standard deviation and the largest median value. GL and GB were the smallest measurements and had the least dispersion (lowest SD values). Male standard deviations were slightly higher than female ones for all predictors except LCL.

Table 5-8 Descriptive statistics for training dataset.

	Male				**Female**			
Predictor	Median	Range	SD	SE	Median	Range	SD	SE
AD	50.08	41.81-58.39	2.86	0.01	44.49	37.98-52.60	2.56	0.01
TAD	47.84	40.51-58.87	2.98	0.01	42.34	38.14-52.99	2.66	0.01
FHD	45.58	38.46-54.59	2.80	0.01	40.20	31.97-47.34	2.48	0.01
THD	45.60	36.14-53.62	2.79	0.01	39.96	31.57-47.49	2.47	0.01
MCL	63.77	52.92-74.09	3.65	0.01	57.19	48.05-66.96	3.58	0.01
BB	79.91	67.27-93.63	4.32	0.02	71.10	60.76-85.58	4.13	0.02
LCL	64.93	52.48-76.38	3.56	0.02	58.67	50.39-70.49	3.76	0.02
TPB	75.22	61.69-87.12	3.77	0.01	66.46	58.96-77.76	3.47	0.02
GL	38.94	31.70-46.11	2.20	0.01	33.66	29.72-41.47	1.91	0.01
GB	27.56	20.40-33.41	2.06	0.01	23.74	19.67-29.21	1.77	0.01
EB	62.39	48.40-71.30	3.57	0.01	54.29	47.52-69.84	3.42	0.01
HHD	44.06	34.86-53.63	3.05	0.01	38.50	24.93-47.46	2.97	0.01

All values given in mm

5.3.4 SEX PREDICTION MODELS

Accuracies and goodness-of-fit for all univariate and multivariate Naïve Bayes, KNN, decision tree, DFA and ensemble algorithms for each predictor or combination of predictors are analysed in detail in Appendix C.

The algorithm used generally had a very minor impact on the *k*-fold goodness-of-fit for any predictor or combination of predictors. The most notable differences between the different algorithm types were in the skewing of accuracies toward one sex. This was more prevalent in univariate models than multivariate models and most pronounced for NB and decision tree models. Across all predictor groupings, discriminant function models tended to present the least sex bias, highest estimated accuracies, and most consistency across cross-validations.

For these reasons, discriminant function analysis was selected as the best algorithm for sex classifier training.

5.3.5.1 OPTIMAL MODEL - TABLE 5-9

The overall best performing algorithm type across all models was DFA. The most useful predictors in multivariate modelling (in order of relevance) were TPB, GL, LCL, EB, THD and TAD (Appendix Figure D1). Discriminant models were thus trained with an increasing number of predictors in order of rank.

The most accurate univariate model, GLdiscr was able to produce a *k*-fold CV accuracy of 88.7%. Adding a second predictor, TPB, increased prediction accuracy to 90.8%. A third predictor, LCL, improved the model once again to yield an accuracy of 92.7%. Adding EB, the fourth predictor, enhanced the accuracy once again, although by a smaller margin, to 93.1%. Adding additional variables was unable to improve the model any further.

The APdiscr model (Appendix Table C4), which included all 12 predictors, achieved a *k*-fold CV accuracy of 92.8%. Thus the 3-predictor model was selected as the optimal model as it best balances maximal prediction accuracy, minimal redundancy, and low material requirement.

Table 5-9 Accuracy and goodness of fit metrics for 'Optimal' sex prediction model.

	Training Accuracy				***K*-fold Accuracy**				**Holdout Accuracy**			
Model	All	Male	Female	Kappa	All	Male	Female	Kappa	All	Male	Female	Kappa
GLdiscr	88.87	88.11	89.60	0.78	88.66	88.11	89.20	0.77	83.33	83.33	83.33	0.67
Model (1)	91.22	90.34	92.04	0.82	90.76	89.37	92.04	0.81	86.86	88.68	85.19	0.74
Model (2)	92.24	92.54	91.96	0.84	92.71	93.03	92.41	0.85	86.50	90.38	83.02	0.73
Model (3)	93.08	92.86	93.27	0.86	93.08	92.86	93.27	0.86	87.47	90.38	84.91	0.75
Model (4)	93.08	92.86	93.27	0.86	93.08	92.86	93.27	0.86	87.47	90.38	84.91	0.75
Model (5)	93.08	92.86	93.27	0.86	93.08	92.86	93.27	0.86	87.47	90.38	84.91	0.75

(1) TPB, GL
(2) LCL, TPB, GL
(3) LCL, TPB, GL, EB
(4) FHD, LCL, TPB, GL, EB
(5) AD, FHD, LCL, TPB, GL, EB
Model selected as ‘Optimal’ model highlighted in **blue**.

5.3.5.2 BEST FIT MODELS

For all predictor combinations, discriminant function models were selected as the best fit models. The accuracies of and discriminant functions for each of the best fit models are presented in Table 5-10 and Table 5-11.

All best fit models are presented in Table 5-11. The most accurate univariate model was GLtree followed by TPBdiscr and EBdiscr. ScapulaDiscr was the most successful multivariate bone model closely followed by HumerusDiscr. None of the multivariate joint models were more successful than the bone models. Scatter plots comparing the known sex of individuals to the predicted sex using particular discriminant models can be found in Appendix C (Figure C2.1a and Figure C3.3). These show that in both 2 and 3 dimensions, discriminant functions struggled to differentiate between individuals in the region of overlap between males and females. APdiscr was marginally more accurate than the 3-predictor Optimal model. The best model, given complexity, redundancy, and accuracy, was the 3-predictor Optimal model.

Table 5-10 Equations for Sex Classification Discriminant functions.

Model	**Discriminant function**
PelvisDiscr	$y = 0.5216(\mathbf{AD}) + 0.3462(\mathbf{TAD}) - 40.4586$
FemurDiscr	$y = 0.1564(\mathbf{FHD}) + 0.3730(\mathbf{THD}) + 0.1151(\mathbf{MCL}) + 0.3019(\mathbf{BB}) - 0.2232(\mathbf{LCL}) - 38.7332$
ScapulaDiscr	$y = 1.1519(\mathbf{GL}) + 1.0250(\mathbf{GB}) - 68.2684$
HumerusDiscr	$y = 0.009(\mathbf{EB})^2 + 0.0873(\mathbf{EB}) - 0.008(\mathbf{HHD})^2 + 0.4463(\mathbf{HHD}) - 31.9755$
HipDiscr	$y = 0.7394(\mathbf{AD}) + 0.6489(\mathbf{TAD}) + 0.7757(\mathbf{FHD}) + 0.7798(\mathbf{THD}) - 131.4185$
KneeDiscr	$y = 0.0794(\mathbf{MCL}) + 0.1366(\mathbf{BB}) - 0.1828(\mathbf{LCL}) + 0.5907(\mathbf{TPB}) - 45.7092$
ShoulderDiscr	$y = 0.9406(\mathbf{GL}) + 0.3208(\mathbf{GB}) + 0.0338(\mathbf{HHD}) - 43.8471$
APdiscr	$y = 0.1831(\mathbf{AD}) - 0.2075(\mathbf{TAD}) + 0.0238(\mathbf{FHD}) + 0.2391(\mathbf{THD}) - 0.0772(\mathbf{MCL}) - 0.0221(\mathbf{BB}) - 0.2683(\mathbf{LCL}) + 0.3540(\mathbf{TPB}) + 0.6735(\mathbf{GL}) + 0.1282(\mathbf{GB}) + 0.1992(\mathbf{EB}) - 0.0729(\mathbf{HHD}) - 54.7762$
Optimal	$y = -0.2349(\mathbf{LCL}) + 0.5387(\mathbf{TPB}) + 0.8067(\mathbf{GL}) - 53.0072$

If y>0 then the individual is classified as a male, if y<0 then the individuals is classified as a female and if y=0 then the sex of the individual is uncertain

Table 5-11 Goodness of fit and univariate sectioning points for all best fit models.

		Training Accuracy				***K*-fold Accuracy**				**Holdout Accuracy**			
Model	Sectioning pt.	All	Male	Female	Kappa	All	Male	Female	Kappa	All	Male	Female	Kappa
ADdiscr	47.49	84.39	83.94	84.82	0.69	84.39	83.94	84.82	0.69	86.14	91.80	80.65	0.72
TADdiscr	45.31	82.61	83.53	81.71	0.65	82.81	83.53	82.10	0.66	85.33	90.16	80.65	0.71
FHDdiscr	43.03	87.12	86.25	87.95	0.74	86.91	86.25	87.55	0.74	82.31	86.44	78.33	0.65
THDdiscr	43.00	86.50	85.83	87.15	0.73	86.71	86.25	87.15	0.73	84.81	91.53	78.33	0.70
MCLdiscr	60.59	80.12	79.08	81.12	0.60	80.33	79.08	81.53	0.61	79.97	81.67	78.33	0.60
BBdiscr	75.70	84.02	84.10	83.94	0.68	83.40	83.68	83.13	0.67	83.30	85.00	81.67	0.67
LCLdiscr	45.38	80.94	79.08	82.73	0.62	80.53	78.66	82.33	0.61	75.88	73.33	78.33	0.52
TPBdiscr	70.97	86.73	85.71	87.66	0.73	87.17	86.64	87.66	0.74	87.39	92.73	82.46	0.75
GLdiscr	36.76	88.87	88.11	89.60	0.78	88.66	88.11	89.20	0.77	83.33	83.33	83.33	0.67
GBdiscr	25.18	84.62	86.48	82.80	0.69	84.62	86.48	82.80	0.69	88.29	91.67	85.00	0.77
EBdiscr	58.20	87.32	87.34	87.30	0.75	87.32	87.34	87.30	0.75	85.15	88.52	81.97	0.70
HHDdiscr	41.62	83.44	82.70	84.13	0.67	83.44	82.70	84.13	0.67	82.66	86.89	78.69	0.66
PelvisDiscr	-	84.58	84.74	84.44	0.69	84.58	83.94	85.21	0.69	86.96	91.80	82.26	0.74
FemurDiscr	-	87.70	86.61	88.76	0.75	87.50	86.19	88.76	0.75	83.96	89.83	78.33	0.68
ScapulaDiscr	-	90.49	90.16	90.80	0.81	90.28	90.16	90.40	0.81	85.82	86.67	85.00	0.72
HumerusDIscr	-	90.18	90.72	89.68	0.80	89.98	90.30	89.68	0.80	88.37	93.44	83.61	0.77
HipDiscr	-	86.13	85.53	86.69	0.72	86.34	85.96	86.69	0.73	83.84	89.66	78.33	0.68
KneeDiscr	-	87.91	87.87	87.95	0.76	87.30	86.61	87.95	0.75	83.13	88.14	78.33	0.66
ShoulderDiscr	-	89.79	88.84	90.69	0.80	89.38	87.98	90.69	0.79	84.19	83.33	85.00	0.68
APlinearDiscr	-	93.49	93.26	93.69	0.87	92.77	92.75	92.79	0.85	85.35	90.20	81.13	0.71
Optimal	-	92.24	92.54	91.96	0.84	92.71	93.03	92.41	0.85	86.50	90.38	83.02	0.73

Values smaller than the sectioning point would indicate that an individual is female whilst values larger than the sectioning point would indicate a male

5.4 *POST HOC* SAMPLE SIZE ANALYSIS

The optimal number of predictors for a discriminant model was confirmed by plotting a curve of model error at different sample sizes. Curves were then plotted to compare relative model error for different machine learning algorithms at different training sample sizes.

5.4.1 NUMBER OF PREDICTORS

Increasing the number of predictors in the model from 1 to 3 caused an obvious decline in mean error at larger sample sizes but, increasing the number of predictors beyond 3 had very little impact on the mean error of the model, especially at larger sample sizes (Figure 5-16). Maximal accuracy and minimal redundancy were thus achieved using a 3-predictor (LCL, GL and TPB) discriminant model. This confirms the findings from section 9.2.1.4 which suggested that the optimal number of predictors is 3.

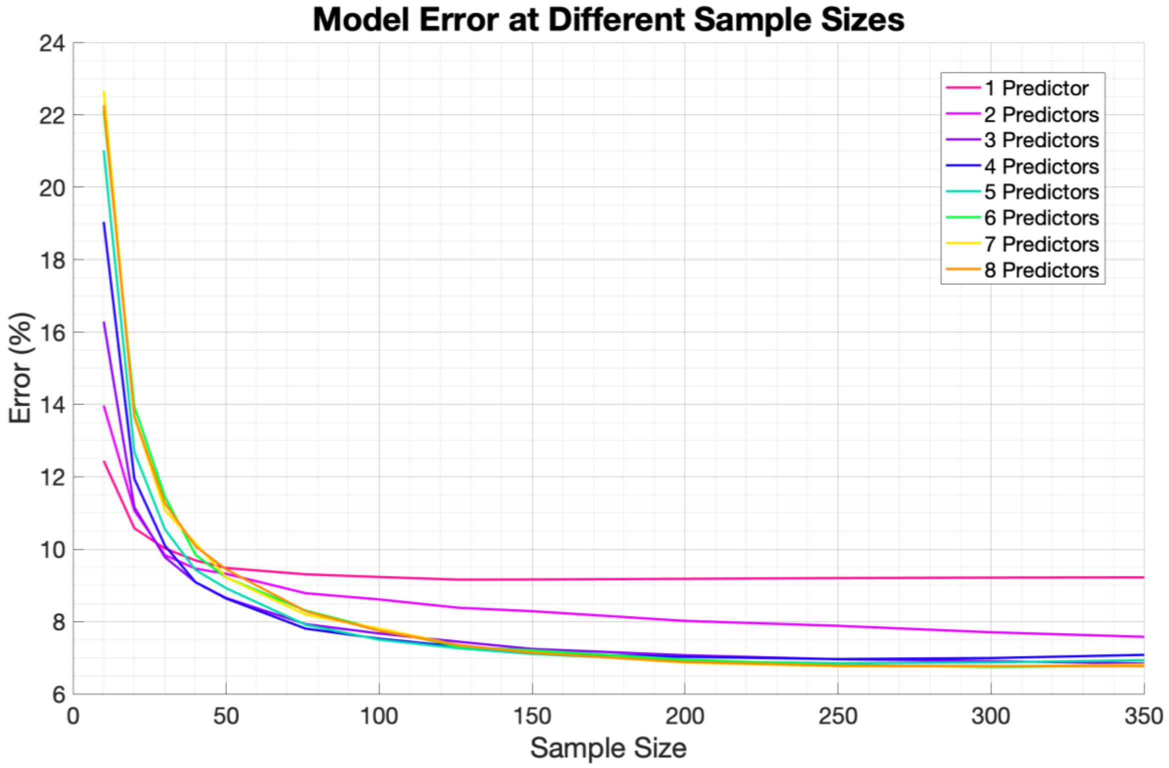


Figure 5-16 Mean model error for linear discriminant models trained 1 000 times with 1 predictor (GL), 2 predictors (GL, TPB), 3 predictors (GL, TPB, LCL), 4 predictors (GL, TPB, LCL, EB), 5 predictors (FHD ,GL, TPB, LCL, EB), 6 predictors (AD, FHD ,GL, TPB, LCL, EB), 7 predictors (AD ,FHD, THD, GL, TPB, LCL, EB) and 8 predictors (AD, FHD, THD, GL, GB, TPB, LCL, EB) at increasingly large sample sizes.

The optimal number of predictors for each algorithm type is explored in Appendix D.

5.4.2 SAMPLE SIZE FOR DIFFERENT ALGORITHMS

Differences in model error at different samples sizes were explored for each algorithm type.

5.4.2.1 LINEAR DISCRIMINANT ANALYSIS - FIGURE 5-17

The mean model error declined rapidly from a sample size of 10 to 100 and then plateaued at a sample size of 250. Since it was not standard practice to train a model 1000 times and select the mean model, we cannot assume that any final model would follow the trend of the mean error line, hence the 95% confidence interval was included. At smaller sample sizes (<50), the 95% CI was very wide meaning that performance estimates were highly variable and dependent on the study sample. Therefore, when a large sample was used to validate the highly variable models trained using small samples, a wide range of accuracies ensued. The width of the CI declined exponentially between a sample size of 10 and 100 and then tended toward a width of zero as sample size continued to increase. The width of the CI at a sample size of 150 was 2.1% compared to 0.6% at a sample size of 400. This means that as sample size grew, the CI shrunk and the certainty that the reported accuracy of a model was correct improved.

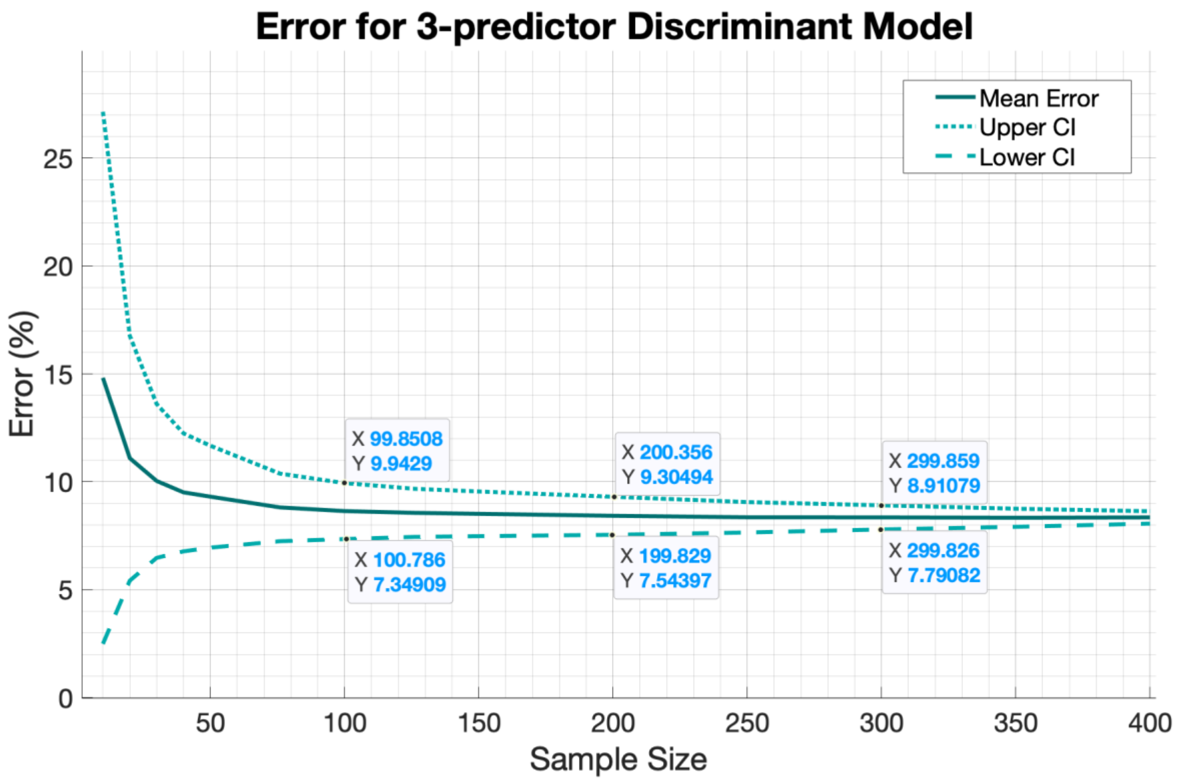


Figure 5-17 Mean classification error and 95% confidence interval (CI) at different sample sizes for a 3 variable discriminant function model trained 1000 times at each sample size.

5.4.2.2 TREE MODELS - FIGURE 5-18

The mean error curve for tree models declined exponentially between a sample size of 10 and 50 and then followed a mostly linear decreasing trend as sample size continued to increase. It is likely that mean error would continue to decline if sample size was increased beyond 400 observations. At smaller sample sizes, the CI was very wide. The difference in error between the upper and lower bounds of the CI was 7% at a sample size of 100. The CI at a sample size of 400 was still wide (~3%), especially when compared to the CI for the discriminant model which was only 0.6% at a sample size of 400. Given this curve, a larger sample size would be needed to make conclusions about the optimal sample size to use when training a decision tree model. A sample of 400 was insufficient to train an optimal, reliable decision tree model to predict sex.

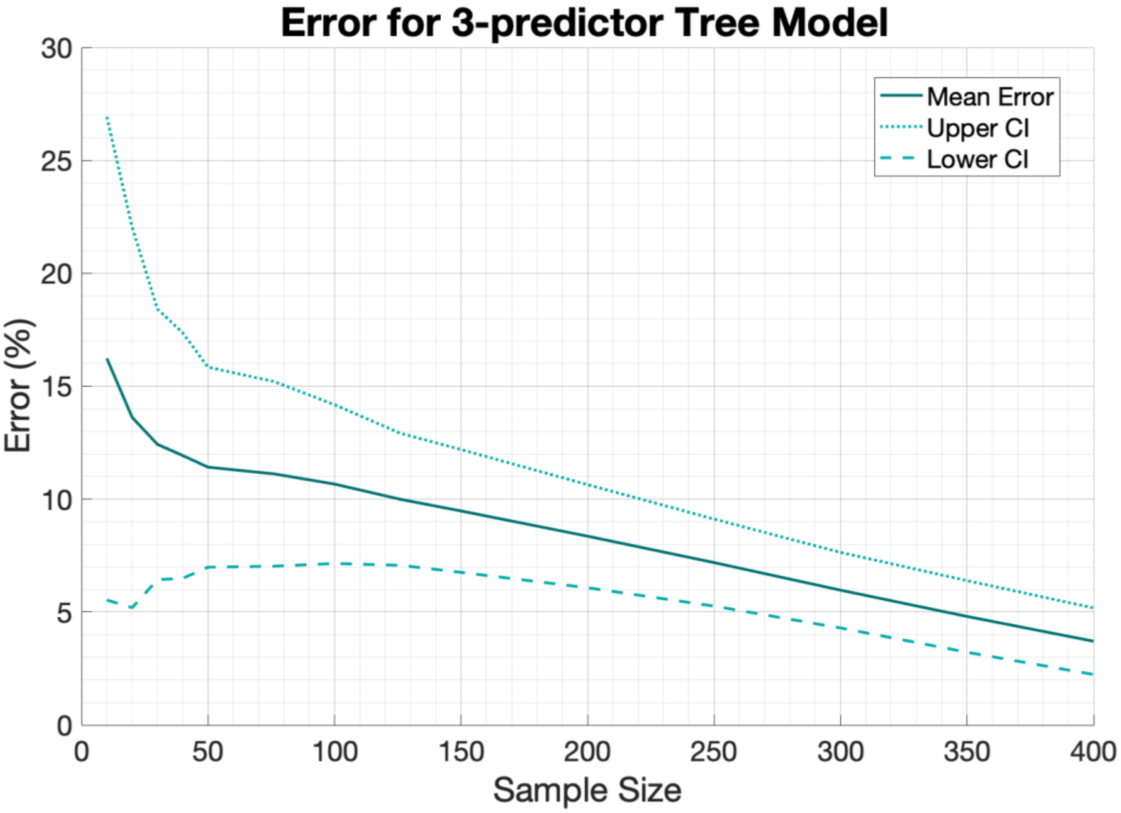


Figure 5-18 Mean classification error and 95% confidence interval (CI) at different sample sizes for a 3 variable decision tree model trained 1000 times at each sample size.

5.4.2.3 ENSEMBLE MODEL - FIGURE 5-19

The sample size versus error curve for models trained using ensemble algorithms followed a similar trend to the decision tree curve. Compared to the tree curve, mean model error for ensemble models was slightly lower and the width of the confidence interval was narrower. The lower bound of the 95% CI reached an error of 0% at a sample size of 400. The width of the CI became narrower as sample size increased thus improving the reliability of accuracy estimates compared to tree models. A larger sample size would be needed to determine the optimal sample size for a 3-predictor ensemble model and the point at which the mean error would plateau. A sample size of 400 observations was insufficient to train an optimally accurate, highly reliable sex predicting ensemble model.

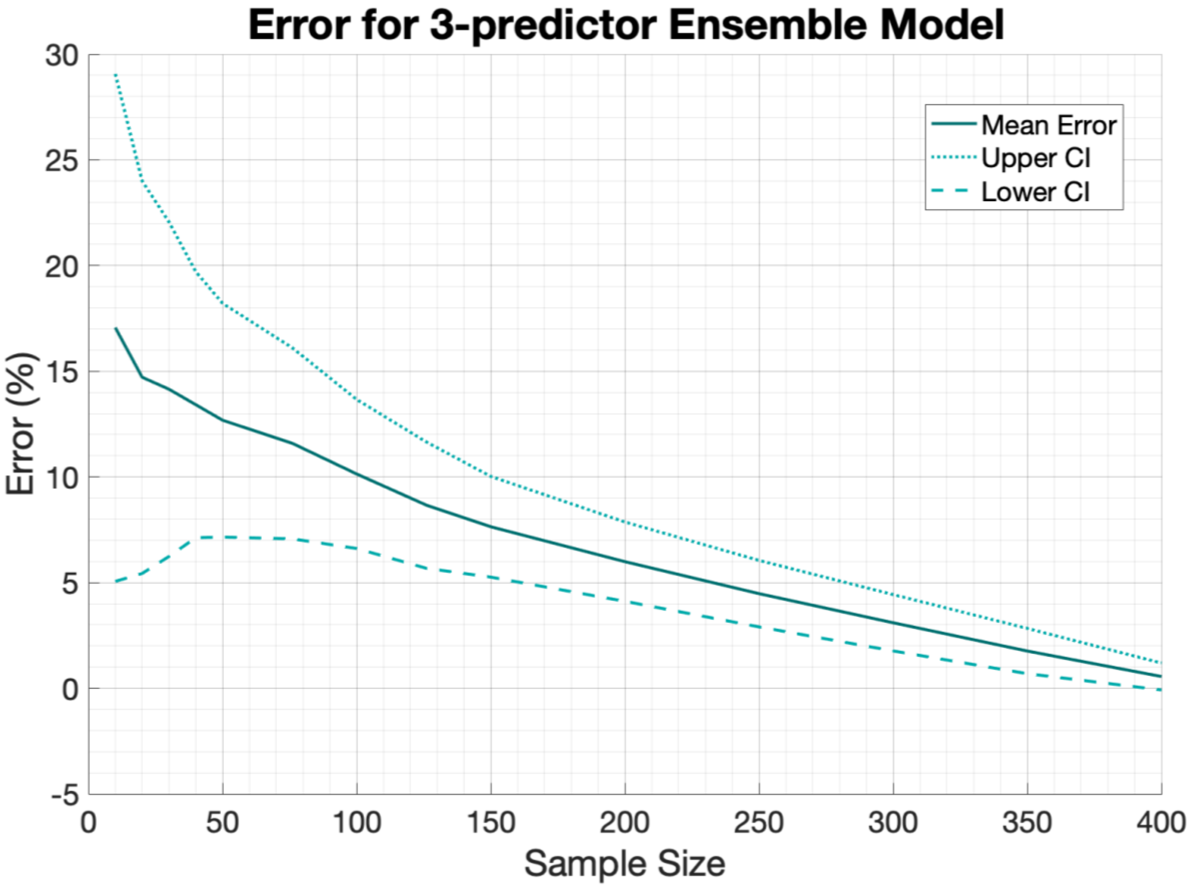


Figure 5-19 Mean classification error and 95% confidence interval (CI) at different sample sizes for a 3 variable ensemble model trained 1000 times at each sample size.

5.4.2.4 NAÏVE BAYES MODEL - FIGURE 5-20

Error for the NB model declined exponentially between a sample size of 10 and 50 but plateaued with a mean error of 10.3% at a sample size of 50. Increasing the sample size beyond 50 individuals did not reduce mean error but did narrow the width of the confidence interval. The CI at 50 Individuals was relatively narrow at only ~2.0% compared to ~5.0% for discriminant models at the same sample size. The confidence interval for NB models was especially narrow at a sample size of 400. This suggested that error estimates for NB models with sample sizes of 400 or more were highly reliable. Despite this, discriminant models were superior to NB models as the mean error for discriminant models plateaued at a smaller error percentage than for NB models therefore making discriminant models more accurate for the given combination of predictors.

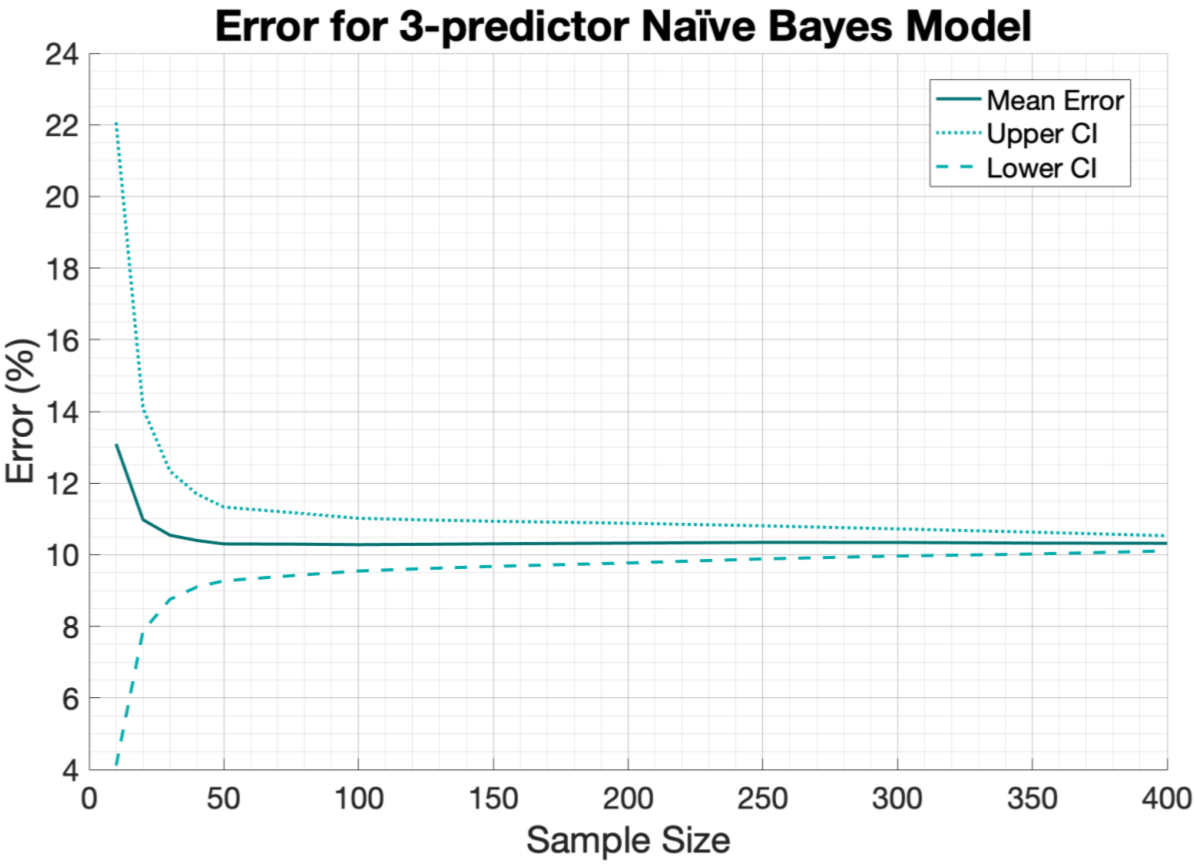


Figure 5-20 Mean classification error and 95% confidence interval (CI) at different sample sizes for a 3 variable naïve Bayes model trained 1000 times at each sample size.

5.4.2.5 KNN MODEL - FIGURE 5-21

Mean model error steadily declined as sample size increased although, it was uncertain whether mean error would continue to decline for sample sizes exceeding 400. The lower bound of the CI tended toward zero at a sample size of 400. This showed that KNN models could estimate sex with 100% accuracy, but it is important to remember that KNN models tend to overfit.

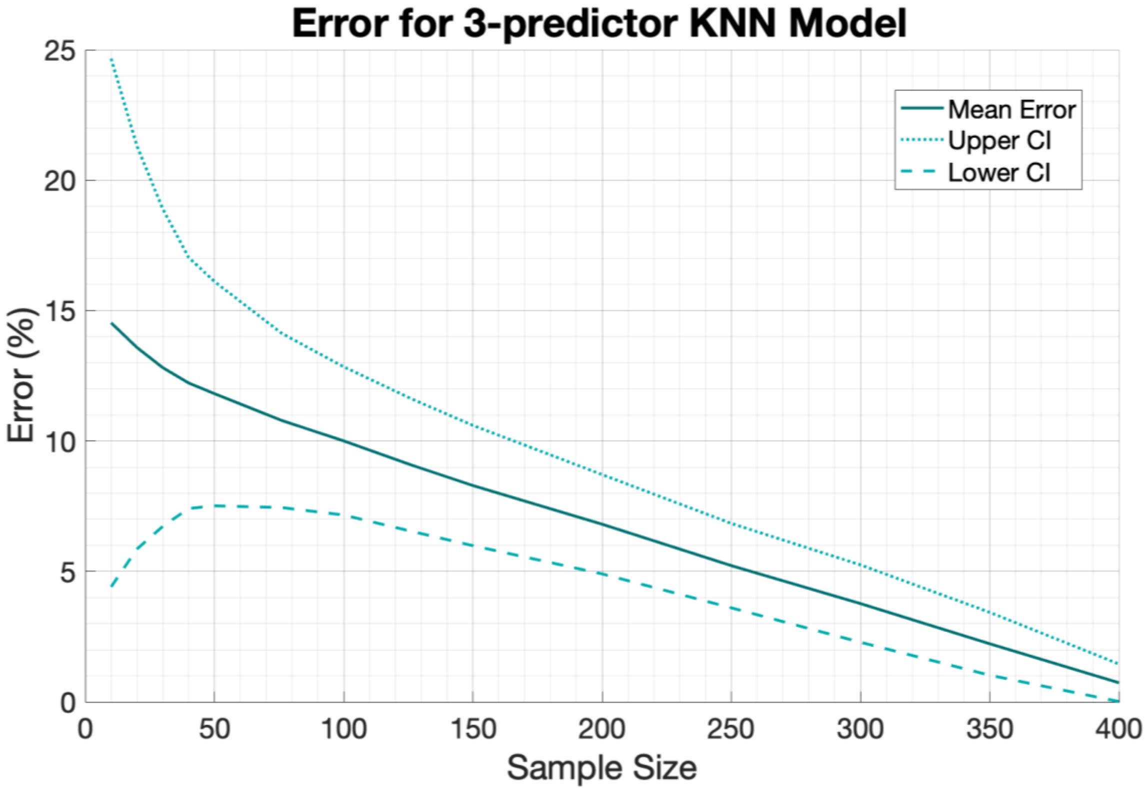


Figure 5-21 Mean classification error and 95% confidence interval at different sample sizes for a 3 variable KNN model trained 1000 times at each sample size.

CHAPTER 6: DISCUSSION

6.1 BASELINE STATISTICAL ANALYSES

Pertinent results of observer error analysis, normality testing, bilateral asymmetry assessment, sexual dimorphism investigation and evaluation of age-based differences will be discussed and explained in the sections to follow.

6.1.1 OBSERVER AGREEMENT

Calculating observer reliability is an important step as it represents the extent to which the collected data are correct representations of the variables measured (McHugh, 2012). The intention of this study was to use existing sexually dimorphic osteometric parameters to provide sex estimation methods which can be applied to the identification of adult human skeletal remains. This relies on it being possible for multiple people to collect data in the form of skeletal measurements. Immediate contention arises as to whether multiple human observers can be consistent in their collection of data. Whilst perfect agreement is not expected, trends where an observer's measurements are always slightly larger or smaller than the baseline, are unsatisfactory.

From the results obtained using Lin's CCC, it was concluded that all 12 of the mensurations in this study were reliable and replicable. An external observer was able to achieve almost identical results and remeasurement by the same observer also yielded very similar results. Thus, given the high CCC scores, none of the mensurations needed to be excluded or revised.

Interestingly, the scores between observers were slightly lower than the within observer scores. These differences were minor given that CCC values for all measurements easily fit into not only the accepted range but also the exemplary range, but it was important to explore these issues and find a potential solution to avoid measurement bias.

The lowest recorded inter-observer CCC value was 0.954 for LHHD. This slightly lower score was likely the result of a single outlying value as shown in Figure 5-1. When the dataset entry for the outlying individual was investigated, the intra-observer and initial observer measurements for the outlying datapoint were very similar for all mensurations except the humeral ones. Oddly, the inter-observer humeral measurements for the right bone were much smaller than the initial observer measurement and vice-vera for the left bone (Table 5-2). This was true for both humeral mensurations. It is impossible to know for sure without repeating the measurement, but the evidence suggests that the left and right bones were switched.

Therefore, incorrect siding of a bone was noted as a potential source of error and considered when deciding whether or not to design side-specific sex estimation models. This will be discussed when bilateral variation is explored in section 6.1.4.

6.1.2 OUTLIERS

A substantial number of interesting outliers existed in the dataset, specifically many extremely small male values. These could possibly be attributed to the underlying demographic of the sample including ancestry, age, and secular trend.

Although data on the ancestry of individuals in the study was not collected, pre-selection of individuals sought to build a demographically representative sample consisting of most SAAA, SAMA and SAEA to avoid demographic biases and overrepresentation of minority groups. The sample set mostly represents SAAA, given that this group forms the majority ancestry group in SA. SAMA, who are a highly variable group, often featuring individuals of shorter stature (Arendse, 2018), represent less than 10% of the sample dataset. It is possible that the small sample of SAMA led to some values in the normal SAMA range being represented as extreme values. This may explain some of the outlying small male and female values. Extreme values like these are an expected feature in highly heterogenous datasets.

The impact of age and secular trend is discussed in section 6.1.5.

The presence of outlying values and their potential effect on means, standard deviations, and possible effect on models (Jackson & Chen, 2004) is acknowledged but given that the outliers are not errors but rather form part of the natural variation in the study population, the decision was taken to retain all datapoints. Any *post hoc* manipulation of data which is certain to increase the chances of finding what we want to find and supporting our thesis is dangerous and has thus been avoided (Aguinis, Gottfredson & Joo, 2013).

6.1.3 NORMALITY TESTING

The jbtest (Thadewald & Büning, 2007) assesses the skewness and kurtosis of a distribution whilst the kstest (Massey, 1951) is a goodness-of-fit test and is highly sensitive to extreme values. Therefore, failure to present the expected symmetry and central tendency of a gaussian curve would cause failure of the normality hypothesis when using the *jbtest* whilst histogram bins which are unexpectedly tall or short compared to a gaussian distribution would result in failing the normality hypothesis when using the *kstest*.

All-but-3 of the mensurations which were tested failed both the *jbtest* and *kstest* and thus convincingly failed to follow a normal distribution. Further visual exploration of the 3 inconclusive mensurations provided insight into why they in one instance failed the normality test but in the other passed.

LMCL (

Figure 5-4) appeared to be multimodal with many bins containing extreme values which did not follow the trend of a standard normal distribution. These extreme values explained why the normality hypothesis was rejected by the *kstest*. The *jbtest* accepted the normality hypothesis because the kurtosis roughly matched a normal distribution and the data were symmetrically distributed, with approximately half of all values falling either side of the mean.

RMCL, in Figure 5-5, interestingly appeared to be quite normally distributed in terms of kurtosis and symmetry – hence why the *jbtest* probably accepted the null hypothesis that the data follow a normal distribution. However, upon closer inspection, in Figure 5-6, where the same data were plotted with more bins, the data were noisy which explained why the *kstest* rejected the normality hypothesis.

RHHD, visualised in Figure 5-7 had an asymmetrical shape which explained why the normality hypothesis was rejected by the *kstest*. When Figure 5-7 was closely inspected, it was difficult to understand how the jbtest for normality gave the most certain result (p = 0.339) out of all the mensurations that the data *were* normally distributed. One tail was visibly fatter than the other making the data asymmetrical, which should have caused the *jbtest* to reject normality.

Given the evidence, none of the mensurations convincingly followed a typical standard normal distribution. As a result of these findings, further statistical testing applied non-parametric testing procedures. Non-parametric alternatives to common parametric tests are provided in Table 6-1 along with examples of when they may be applied.

Table 6-1 Analogous parametric and nonparametric procedures.

Analysis Type	**Example**	**Parametric Procedure**	**Nonparametric Procedure**
Compare two independent groups	Sexual Dimorphism (Male vs. Female)	Two-sample t-test	Wilcoxon rank-sum tests
Compare more than two independent groups	Age-related differences (Child, Adolescent, Adult, Advanced Age)	Analysis of Variance (ANOVA)	Kruskal-Wallis's test
Compare two measurements from the same individual	Bilateral Variation (Left vs. Right)	Paired t-test	Wilcoxon signed-rank test
Estimate the degree of association between two variables	Observer Error (Observer 1 vs. 2)	Pearson's *r*	Spearman's rank correlation

6.1.4 BILATERAL ASYMMETRY DISCUSSION

The results of Wilcoxon Signed Rank testing in this study concluded that there was no significant bilateral variation between paired left- and right-sided measurements for any mensurations except GB (Table 5-4). Additionally, differences between median values for left- and right-sided measurements were, in most cases, smaller than or very similar to standard error showing that there was very little observable difference between left- and right-sided bones in the sample.

The Kanchan et al. (2008) hypothesis that the upper dominant limb would be more robust, and the lower limbs would present less asymmetry was not met. The right-sided measurement was on average larger for both upper and lower limbs (rather than the contralateral lower limb being larger), and the majority (60.0-90.6%) of right-sided measurements fell within 1mm of their paired left measurement. During data collection, a 1mm margin of error was acceptable between measurements in the series of 3 measurement repeats so a difference of 1mm between paired left- and right-sided measurements was acceptable.

The most variable mensurations were further explored in violin plots which visualised the probability density of measurement values. These plots showed that whilst there was slight variation in the distribution of the data, the similarities were overwhelming and differences between left- and right-sided measurements were minor.

Given these findings, left- and right-sided bones did not differ significantly on either a pair-by-pair basis nor a sample wide basis. As a result, developing separate models for left- and right-sided bones seemed unnecessary as the differences were too minor for the models to have significantly increased discriminatory power. Additionally, needing to side a bone before classifying sex added an extra dimension for potential error. This became clear during inter-observer analysis where instances of the left- and right- humeri being misidentified arose. These findings established a firm basis for the exclusion of left- and right-side specific models.

Dabbs & Moore-Jansen (2010) used each bone, regardless of side, as a separate observation to maximise the number of bones in their study. The concern with this methodology was that certain individuals (where both bones were available) would be double represented, and thus bore the risk of overrepresenting extreme and outlying individuals within the dataset.

Another option was to use mean values only (taken as an average of the left and right measurements). This, however, would have led to needing to discard a substantial amount of data as many of the skeletons in the sample only had either the left- or right-sided bone available.

One study found 10 of their 54 mensurations to be significantly asymmetrical and, in cases where there was no asymmetry, chose to use only one bone with the right bone taking preference (Jerković et al., 2020). Given that the decision on how to manage left- and right-sided bones in this study took place after data collection, this approach would have led to a large amount of data being discarded.

Therefore - reiterating that bilateral asymmetry was minor in this study sample - to maximise the available data without double representing any individuals, a mean value was calculated and used whenever both the left- and right- measurement were available and when only either the left or right bone was available, the available measurement was used alone.

6.1.5 SEXUAL DIMORPHISM

For all 12 mensurations in this study, male measurements were significantly larger than female measurements ($p<0.05$ for *ranksum*). Owing to these results, none of the mensurations were excluded when building sex classification models.

The most sexually dimorphic mensurations were the 2 smallest ones in this study, GL, and GB. These are both measurements taken from the glenoid fossa of the scapula. Despite having the smallest measurement values, these mensurations also had the smallest variance thus suggesting that the glenoid fossa was truly the most dimorphic element studied. The scapula was found to be a later-growing skeletal element which continued to grow and develop after adolescence and this later-growing nature was linked to greater levels of sexual dimorphism (Humphrey, 1998).

For GL there was very good separation between male and female medians and only a small degree of overlap between male and female values. These results indicated that GL would be the most useful predictor variable for sex estimation. The second most dimorphic mensuration based on Wilcoxon Signed Rank testing was GB. The dimorphism in GB was less visually striking than expected given the large p-value and GB is likely to be a less accurate predictor of sex than GL due to the larger region of overlap between male and female values. In addition to their Wilcoxon Signed Ranks scores, GL and GB also had the greatest percentage difference between male and female medians and the least variance.

GL and GB are good candidates for building univariate and multivariate sex classification models and are likely to produce the most accurate models.

The least dimorphic measurements were MCL and LCL given that they had the highest p-values (although still very small). The medians for these mensurations were well-separated although they had the smallest percentage difference between males and females and the variance was large (Table 5-4). In the bar charts for both mensurations (Figure 5-9 and 10), the long tails for male and female measurements caused the large margin of overlap between male and female values. Whilst MCL and LCL were the least dimorphic mensurations in this study, they did present clear distinction between groups. These mensurations would be expected to be least effective in univariate analysis but possibly more useful when combined with other predictors for multivariate techniques. Other studies which assessed 'Black' and 'White' Americans have also found upper limb measurements from the scapula and humerus to be more dimorphic than measurements from the distal femur (Spradley & Jantz, 2011).

6.1.6 AGE AND TIME

Sample Demographic

Every effort was made during data collection to ensure that the study cohort was not age nor age-ancestry biased despite the biases presented by skeletal collections (Maass & Friedling, in 2019, Dayal et al., 2009). The UCT skeletal repository is especially biased towards older SAEA individuals, which bears the risk of overrepresenting a historically socio-politically favoured minority group (da Silva, 2006).

The demographic data presented in the results Figure 5-14 and

Figure 5-13 showed that neither the distribution of individuals by age-at-death nor year-of-birth presented concerning patterns. There was a clustering of individuals within the study sample between with an age-at-death of between 40 and 60 years. This correlated well with the uptick in deaths seen in the South African population around the age of 40 (Statistics South Africa, 2018) whilst maintaining forensic relevance for individuals who died earlier in life, rather than focussing only on deaths more associated with old age.

Secular Trend

Evidence has been presented to suggest that there was a weak positive secular trend in both male and female SA stature over the period of 1880 to 1990 (Henneberg & van den Berg, 1990). None of the measurements in this study were bone lengths, as the focus was placed on joint dimensions, thus, it was not possible to directly compare secular trend in joint dimensions in this study to trends in stature reported by Henneberg & van den Berg (1990)

There was no correlation between year-of-birth and female measurement value (Table 5-7). therefore, there was no significant change in the joint dimensions of the female individuals in this study in response to a more recent year of birth. Secular trend therefore did not influence female measurements in this study.

For all male mensurations, except LCL, there was a significant but weak negative linear correlation between measurement value and year of birth which signified a decreasing trend in joint dimensions in response to a more recent birth year.

Klales (2016) punctuated fears that secular trend may increase the difficulty of sex estimation but also provided evidence to show that secular trend had led to increased sexual dimorphism in pelvic morphology. Although the impact of secular trend was small for females in this study sample, the negative trend in males may have caused an increased overlap between male and female measurement values and thus made separation of the sexes more challenging.

Given these findings, it would be advisable to limit the year-of-birth range as much as possible to reduce the possible influence of secular trend on skeletal data and to have a study sample which is most representative of the contemporary living population. Due to limitations in available skeletal material, the year-of-birth range for this sample spanned more than 100 years but including individuals from more skeletal collections including the Kirsten and UP bone collections could help to diminish these limitations (L'Abbe, Loots & Meiring, 2005; Alblas, Greyling & Geldenhuys, 2018).

Bone Changes with Advancing Age

There was a weak positive correlation between age-at-death and bone measurement value for all mensurations in males and females (Table 5-7) This can be explained by the underlying mechanisms associated with aging in males and females as described in section 2.3.

Due to correlation between measurement value and age-at-death, older males and females had, on average, larger measurement values. These larger values caused older female measurements to overlap with data from smaller males. Whilst larger values from older males would not hinder separation between the sexes, they would impact means and standard deviations for the dataset.

Vance (2007) assessed age-based differences in SA skeletal remains and found that there was little change in postcranial measurement values for SAAA females but a significant increase in measurement value for SAEA individuals. These incongruous patterns across different ancestry groups may have caused sex estimation accuracies in older SAEA females to be lower than for other groups especially given that older SAEA individuals only make up a very small proportion of the study sample. It was not possible to test this theory given that ancestry data was not collected.

Dividing the dataset into age-based groups like young, middle aged and mature has the potential to improve sex prediction accuracy by shrinking the region of overlap caused by large female values from older individuals. The concern with this kind of methodology is that classification methods using these age segregations would require the predetermination of age before sex could be estimated. Age estimation is not 100% accurate (Franklin, 2010) and thus, the inclusion of age as a prerequisite for sex estimation would introduce added error.

Therefore, whilst dividing the sample into age-specific groups may be able to improve the power of sex classification models, the improved accuracy would need to be very large to justify the compounding error introduced by adding age estimation to the protocol for sex estimation.

6.2 SOURCES OF VARIANCE AND ERROR

Many sources of variance exist when training machine learning algorithms. A discussion on factors which influence replicability and reproducibility including differences caused by learning algorithm, platform and training data can be found in Appendix B. The following subsections will explore the effects of sample size, sex bias and overfitting.

6.2.1 *POST HOC* SAMPLE SIZE ANALYSIS

Post hoc sample size analyses were performed to determine the optimal number of predictors and the optimal sample size for different algorithm types.

6.2.1.1 OPTIMAL NUMBER OF PREDICTORS

Analysis to determine the optimal number of predictors for a discriminant model showed that increasing the number of predictors beyond 3 had very little impact on the resultant error rates if the sample size exceeded 100 observations. This is likely because all of the predictor variables in this study shared a very similar positive linear relationship with one another and sex. Thus, the addition of more than 3 predictors provided little assistance in separating the sexes.

An analysis was also performed to determine the optimal sample size for other algorithm types (Appendix D). Results varied widely however the optimal number of predictors for KNN, and Tree models was also 3 as the inclusion of more predictors did not drastically improve the accuracy of models. KNN and Tree models both rely on clustering of datapoints from the 2 classes. Whilst male and female datapoints did tend to cluster together, there was a fairly large region of overlap between them. The addition of extra predictors refines the boundary between the 2 clusters but given that the skeletal remains of males and females are not discrete but rather exist on a continuum (Loth & Henneberg, 1996), it is not possible to achieve perfect separation of classes especially when using predictors which are so closely related.

6.2.1.2 OPTIMAL SAMPLE SIZE FOR A 3-PREDICTOR MODEL

Given the findings of analysis for the optimal number of predictors, it was decided that sample size analysis would be performed for a hypothetical 3-predictor model for each algorithm type. The error curve and confidence interval for 3-predictor models were explored for each algorithm type.

6.2.1.2.1 Discriminant Models

As the sample size grew and the CI narrowed, the sample encompassed more of the variation in the population and thus, once the full spectrum of variation was accounted for, sex estimation models and their accuracies became more reliable. At larger sample sizes, the final model was less variable because almost all of the variation in the population was accounted for whilst at smaller sample sizes, outlying datapoints were more influential on the final model. So, as the sample size grew and the impact of a single datapoint declined, that the confidence interval became increasingly narrow, and error became model-agnostic. This shows that, as predicted, relatively larger sample sizes led to more consistent performance estimates which were resistant to small variations in the dataset.

There is thus a trade-off between effective gains accumulated by increasing sample size and the 'costs' associated with increasing sample size when training discriminant function models. A sample size larger than the cohort of this study may help to reduce variance and aid in making the final model and prediction accuracy more reliable and representative of the population but is unlikely to considerably improve the mean prediction accuracy of a 3-predictor discriminant model given that increasing the sample size from 250 to 400 had no impact on mean error.

6.2.1.2.2 Tree and Ensemble Models

When 3-predictor decision tree and bagged ensemble tree models were trained at increasingly large sample sizes, the pattern differed considerably from the discriminant function curve.

It was stated previously that decision trees are high variance classifiers which require large sample sizes to reduce variance and, where the sample size is insufficient, ensemble methods can be used to counteract variance. It was clear from the decision tree and bagged ensemble tree curves that this statement translated from theory to practice. The confidence interval for the decision tree curve was very wide, even at larger sample sizes and a bigger dataset would have been needed to reach the optimal accuracy which could be achieved using this algorithm. In comparison, the curve for bagged ensemble trees had slightly lower error but more importantly, a narrower confidence interval. This showed that ensemble algorithms were especially useful to achieve reliable and accurate predictions when the sample size was limited.

Neither the decision tree nor the ensemble model curve for mean error reached a plateau. Mean error continued to decline beyond the maximum tested sample size. This suggested that decision trees and ensemble models may find features in the data which discriminant models cannot and may have the potential to be more accurate and, more reliable than discriminant models, when the sample size is larger than the maximum sample size investigated in this study.

6.2.1.2.3 Naïve Bayes

The 3-predictor NB model curve had a similar shape and trend to the DFA curve. The NB curve reached its minimum mean error (10.30%) a sample size of only 50 compared to the minimum mean error of 8.34% at a sample size closer to 150 for DFA.

The benefits of NB, compared to DFA models, were that the confidence interval was narrower, and the mean error plateaued earlier. The downside to NB was that curve plateaued at a higher percentage error than for DFA models.

If a study were to be performed using a very small sample of between 50 and 100 individuals, then NB would be a good choice of algorithm. In any situation where the maximum potential sample size is less than 100, the NB algorithm would be a good choice because at these sample sizes, the mean error was lower, and the CI was narrower than for DFA models. But, at sample sizes larger than 100, the NB confidence interval was not comparatively narrow enough to justify its use over more accurate discriminant models.

6.2.1.2.4 KNN

Results showed that KNN models could achieve very low mean error rates at small sample sizes however, the mean error curve had not yet plateaued at this study's maximum sample size and, the confidence interval was relatively wide. The curve was plotted using re-substitution of the training data and as such, the mean error presented in the sample size curve for KNN was likely highly influenced by overfitting. Multivariate KNN models were often shown to overfit when exploring the results of modelling in this study.

KNN models are powerful when the data are non-linearly clustered in multiple dimensions (Guo et al., 2003). The data in this study all shared a positive linear correlation with sex. KNN models trained using transformed data which is negatively correlated or differently clustered may improve KNN's ability to discriminate between the sexes.

These findings strongly suggested that, with the available dataset in this study, the KNN algorithm was not the most reliable algorithm option for sex estimation. KNN algorithms would require a much larger sample size to reach their maximal accuracy and minimal variance whilst also mitigating overfitting.

6.2.2 SEX BIAS

Biases are easily introduced into algorithms, especially when the data are not balanced, or one dataset is more variable than another. This is part of the reason why statistical analyses were performed prior to algorithm training. Factors like age-at-death and secular trend were assessed to ensure that sex was not being confounded with other features of the dataset.

A machine learning algorithm will attempt to solve the problem of sex classification by finding the model which most accurately predicts sex. The problem is that, as shown in Figure 6-1, the sacrifice of equal accuracy for both classes may lead to a significantly higher overall model accuracy.

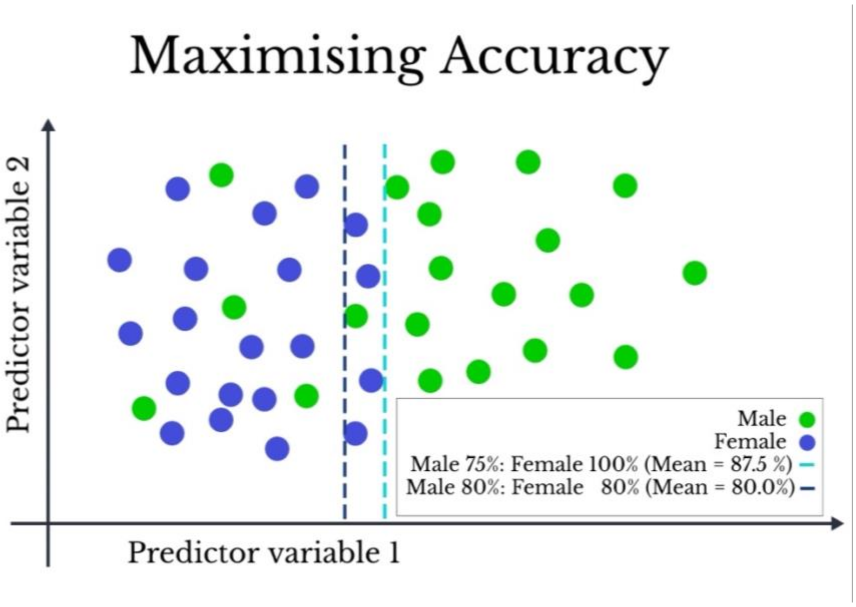


Figure 6-1 Diagram illustrating an example dataset where maximising overall accuracy of classification comes at the cost of one class having a higher classification accuracy than the other.

Higher accuracy for one class is generally caused by the data from that class being more variable and thus more scattered than the other. One class in the dataset being more variable is a feature of the dataset rather than a factor which can be controlled. Thus, if the estimated accuracy for one sex if higher, it isn't necessarily a failure of the algorithm itself, since the algorithm has the primary goal to train a model which predicts sex with the highest possible accuracy. The main problem with a model containing a significant sex bias is that it is inequitable towards one sex.

Male and female accuracies should be equal - or at least close to equal. Given that, according to government statistics, 48.8% of the South African population is male whilst the remaining 51.2% is female (Statistics South Africa, 2019), there would be little benefit in favouring one class over the other. Across most models and variables in this study, there was a slight skewing towards more accurately predicting sex in females than males. This was most prevalent in univariate models for the lower limb.

Some of the univariate tree models showed large differences in accuracy between males and females. TADtree had a large difference of 12.4% between male and female *k*-fold CV accuracies as did ADtree (10.0%), BBtree (21.1%) and LCLtree (17.6%). These sex biases were also present in training accuracies. The sex biases in tree models can be attributed to tree models being especially susceptible to biases due to the nature of the decision tree algorithm (Dietterich & Kong, 1995). Decision tree algorithms find a sectioning point by assigning weights based on 'impurity' (Breiman et al., 1984). Impurity means that the sectioned data contains observations from the other class. So, the algorithm finds the point that will divide the data into its classes in a manner which minimises impurity. If the data for one class is more variable than the other, then datapoints from that class are likely to be scattered amongst the more clustered data from the other class. The more clustered class data will represent a higher proportion of data in the region of overlap and thus the sectioning point will be positioned in a manner which favours the correct classification of less variable class because impurity will be lower (Dietterich & Kong, 1995).

The most biased univariate predictors - AD, TAD, BB and LCL - were all lower limb mensurations. The prevalence of a sex bias in these mensurations could perhaps, in part, be credited to lower limb morphology being more variable in males than in females (Walker, 2005; Pretorius, Steyn & Scholtz, 2006).

Many research outputs in South Africa and abroad have reported discriminant functions and accuracies which are skewed towards a sex for both the upper and lower limbs. Steyn & Işcan (1999) published discriminant functions for epicondylar breadth, where female accuracies for SAAA (91.1%) and SAEA (95.8%) were much higher than male accuracies (86% and 83.6% respectively). Similar trends followed for AD, FHD and HHD (Steyn & Işcan, 1997, 1999; Patriquin, Steyn & Loth, 2005). Even recent international machine learning studies have published skewed results like Curate, Umbelino, et al. (2017) who reported FHD (vhdl) accuracies of 86% for Portuguese females but only 76% for males.

Other papers have shown that it is possible to get balanced and high accuracies using both upper and lower limb measurements. MacAluso (2011) reported discriminant function prediction accuracies of 86.7% for both males and females when using GL and Bubalo et al. (2019) reported 88% accuracy for males and females when using AD.

When selecting best fit functions, those with the least sex bias were preferential when 2 models had similar overall accuracy values. For example, BBdiscr with a *k*-fold kappa value of 0.67 was favoured over BBtree with a higher *k*-fold kappa value of 0.69 because BBdiscr had a sex bias of only 0.6% compared to massive 21.1% for BBtree.

Of the 5 types of machine learning algorithm used to train models, discriminant function analysis resulted in models with the least sex bias whilst maintaining high model accuracies. This is probably the result of a combination of sample size deficiencies when using certain algorithms and the differential influence of extreme values on algorithm training.

6.2.3 OVERFITTING

Overfitting is the phenomena whereby a model or modelling procedure that includes more predictors or coefficients than are necessary or uses more complicated approaches than are necessary is used to solve a machine learning problem (Hawkins, 2004).

The main consequence of overfitting is training a model that is so closely aligned to the training dataset that it is unable to generalise and respond to new data. A visual example of overfitting is presented in Figure 6-2. A perfect, although complex, prediction model is trained which correctly maps all the datapoints in the training set to their respective outputs. When new data are applied to the prediction model, the model is unable to generalise and correctly predict sex for many of the datapoints because the model was too closely aligned to the training data. As shown in Figure 6-2c, a more simplistic model, although less accurate on the training data, would have generalised better to new data.

No obvious overfitting was prevalent in univariate, bone or joint NB, tree, discriminant, and ensemble models within this study however, KNN models, especially multivariate ones, were susceptible to overfitting.

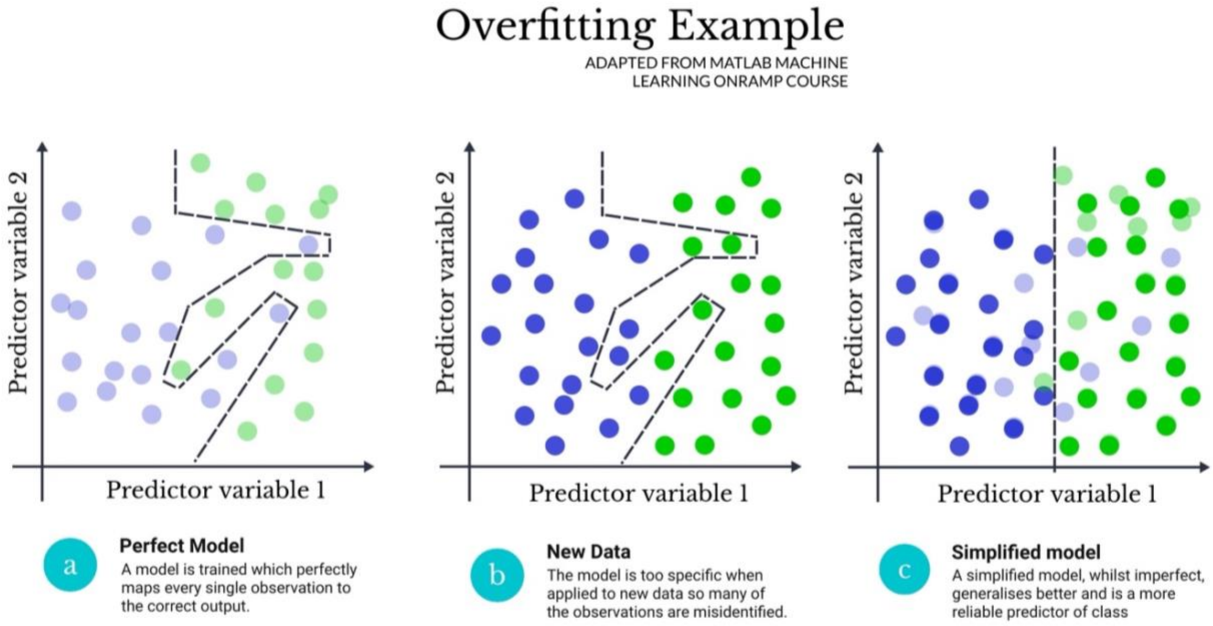

Overfitting Example
ADAPTED FROM MATLAB MACHINE LEARNING ONRAMP COURSE
Predictor variable 2
Predictor variable 1
a
Perfect Model
A model is trained which perfectly maps every single observation to the correct output.
b
New Data
The model is too specific when applied to new data so many of the observations are misidentified.
c
Simplified model
A simplified model, whilst imperfect, generalises better and is a more reliable predictor of class

Figure 6-2 Visual example of model overfitting adapted from the MATLAB machine learning onramp course.

The PelvisKNN model had 100% training accuracy but only 85.38% accuracy when tested using k-fold CV. Similar patterns followed for HipKNN, KneeKNN and ShoulderKNN. Sample size analysis suggested that the number of observations in the dataset for this study was insufficient to train reliable KNN models. Overfitting for the multivariate KNN models was thus likely the result of limited sample size. Consequently, given the limitations of this study, KNN models were not selected as best fit models or recommended for use when attempting to estimate sex.

A common source of overfitting is the inclusion of too many features or predictor variables. The inclusion of many predictors increases the chance of redundancy in features and the inclusion of predictors which are not related to the response variable. Sexual dimorphism was assessed for the predictors in this study and all of them were convincingly related to sex. The inclusion of too many features still posed the risk of feature redundancy in prediction models. The APlinearDiscr models used the largest number of predictor variables (12) and was therefore at the highest risk of overfitting. Predictor contributions to the model were presented in Figure C4b and showed that all predictors did contribute but BB made a very small contribution relative to other predictors. The 3-predictor Optimal model was able to achieve almost the same level of accuracy as the 12-predictor APlinearDiscr model thus suggesting that too many features were included in the APlinearDiscr model.

6.3 SELECTING BEST FIT MODELS

6.3.1 BINARY DECISION TREES

By ease of applicability, decision trees are the most simplistic and easy (Kotsiantis, 2007) to apply models as they require no mathematical calculation and can easily be used in the field without the need for computational software of any sort. Decision trees are useful for quick analysis and results can be verified later in a lab setting using more complex, and likely more reliable, methods if necessary.

Unfortunately, their simplicity means that decision tree models have the least discriminatory power and identify the least complex relationships between predictors. They only use simple sectioning points to divide the data into classes and have few optimisable hyperparameters. Because of this simplistic nature, a small change in the training data may significantly alter the consequent model which is trained by the algorithm and thus change resultant predictions, especially if the sample size is 'small' (Breiman et al., 1984; Dietterich & Kong, 1995).

Findings of sample size analysis suggest that a larger dataset than what was available for model training in this study would be needed to train optimally accurate and reliable decision tree models. These findings insinuate that given the sample size and predictors investigated, decision trees were not ideal predictors of sex.

Another unsavoury feature of decision trees trained in this study was their tendency toward containing a sex bias. In particular, the BBtree model had a difference of 21.14% between male and female *k*-fold CV accuracies. Even multivariate tree models failed to eliminate sex biases. Sex bias for most model types was small when the number of predictors was large however, *k*-fold CV accuracy for males and females using the 12-predictor APtree model had a 11.31% difference. Another study using decision tree models to estimate sex from Portuguese skeletal remains also reported higher sex estimation accuracies for females than males (Navega et al., 2015).

A feature of multivariate decision tree models in this study was that the decision tree algorithm would often discard predictors in favour of a simpler model. The 3 joint-specific decision tree models had this feature. Despite being trained using multiple predictors, only a single node was selected in each of the models. GLtree & ShoulderTree, TPBtree & KneeTree and, THDtree & HipTree are thus homologous model pairs as despite different training data, each pair shares the same sectioning points and model structure.

But why did the decision tree algorithm train Shoulder-, Knee- and HipTree to contain only 1 node? Decision tree algorithms tend to consider small trees with few nodes before they consider larger trees with many nodes (Dietterich & Kong, 1995). If a small tree is grown that can classify the data, then a larger tree will not be considered. Decision tree algorithms select to discard additional redundant predictors in favour of retaining the simplicity of a model. This is a great feature of decision trees as it makes overfitting highly unlikely.

Sample size analysis showed the study sample to be insufficient to train reliable and accurate decision trees. Due to their relative weakness as reliable and equitable predictors at this study's training sample size, decision trees were not selected as best fit models in any cases and were not recommended for use in building models based on metric bone mensurations unless a much larger sample were to be available.

A strategy to improve the classification power of decision trees in the absence of a larger dataset would be to use more homogenous sample data. Population specificity has been widely studied with varying conclusions (Bidmos & Dayal, 2004; Steyn & Patriquin, 2009 Dillon, 2014; Kotěrová et al., 2017). A dataset sourced from individuals from a more localised geographical region or, who died over a more similar division of time with a narrower age-at-death range or, hail from similar socioeconomic circumstances may improve the power of a model however, increasing the specificity of a model would limit its forensic applicability and relevance. Therefore, increasing the specificity of the dataset to improve the prediction power of a model is not recommended.

6.3.2 ENSEMBLE METHODS

Ensemble methods are often used to reduce the model variance associated with decision trees (González et al., 2020). The downside of ensemble methods is that they require a computer and software to run and, these models are difficult to communicate. The benefit of ensemble classifiers is that their increased complexity and larger number of optimisable hyperparameters allow them to find more complex relationships between predictors. This theoretically means that ensemble models should be more accurate to use in a lab environment when the requisite software and equipment are available. Ensemble algorithms would however only be considered superior to any of the other algorithms explored in this study if their models were considerably more accurate.

When comparing relative accuracies of ensemble models to decision tree models, ensemble models were, in most cases, equally accurate. Only 2 ensemble models, TibiaEnsemble and FemurEnsemble outperformed their relative decision tree model. In both cases the improvement in training and *k*-fold CV accuracy from the decision tree model to the ensemble model was less than 2%.

The improvement in accuracy from FemurTree to FemurEnsemble is easily explained. FemurTree, although trained using 5 predictors, had only 1 node. When the ensemble of trees was trained for FemurEnsemble, some of the trees had more than one node thus increasing their complexity and improving their predictive power when compared to the tree model which only used one node.

Despite their theoretical capacity to outperform more simplistic models, the ensemble models trained in this study were never sufficiently better than more simplistic, easier to use models to justify their use. As a result of their inferior accuracy in relation to model complexity, ensemble models have not been selected as best fit models for any predictors or combinations of predictors and are not recommended for use as sex predictors using the available dataset.

6.3.3 DISCRIMINANT FUNCTION ANALYSIS

Discriminant functions are relatively easy to use in the field, as calculations can be performed manually, and are widely used in biological anthropology (Šlaus et al., in 2013 Asala, Bidmos & Dayal, 2004; James MacAluso, 2010; Mokoena et al., 2019; Bidmos & Mazengenya, 2021) as the gold standard although there is little available justification for why to be found in the literature

Coelho & Curate (2019) analysed various machine learning algorithms and concluded that amongst the algorithms shared between this study and theirs, linear discriminant analysis was the best for sex classification.

Analysis of model error when compared to sample size showed that the size of the study dataset was sufficient to yield models with the lowest possible mean error and, minimal variance between models using slightly different datasets. Given the sample sizes available from skeletal collections in SA (L'Abbe, Loots & Meiring, 2005; Dayal et al., 2009; Alblas, Greyling & Geldenhuys, 2018; Maass & Friedling, 2019), DFA is a reliable option.

Whilst binary decision trees rely on a single coefficient or sectioning point to divide data, discriminant functions have both a slope and an intercept coefficient at minimum thus making them more complex in structure than decision trees and theoretically better at discerning between classes (Paluszek & Thomas, 2017). Discriminant functions were selected as the best fit models for all univariate and multivariate combinations of predictors. Whilst other models did, in some cases, have slightly higher *k*-fold CV accuracy estimates - taking into consideration model complexity, ease of use and sex biases - discriminant models were always superior.

Since DFA was the most successful of all algorithm types, an Optimal model was trained using discriminant function analysis. *K*-fold CV accuracy for the 3-predictor Optimal model was almost identical to the estimated accuracy of far more complex the 12-predictor AP model. The 12-predictor model was overcomplicated and thus overfit since the level of complexity was unnecessary. The final Optimal model is recommended above all other models in this study if the 3 mensurations needed for it are available.

Due to model simplicity, ease of use, sex prediction equity, low sample size requirements, high prediction accuracies and minimal between-sample variance - discriminant function analysis was the best machine learning algorithm of all those tested for use in sex estimation from the joint surfaces of long bones.

Therefore, the continued use of discriminant function analysis to derive sex estimation methods in the field of biological anthropology, especially when sample size is between 100 and 500 individuals, is recommended.

6.3.4 NAÏVE BAYES

Naïve Bayes (NB) utilises conditional probability to make predictions of class (MathWorks, 2021e). The NB algorithm is simple and has few parameters thus making it a high bias, low variance classifier which theoretically performs well on a limited dataset (Kotsiantis, 2007). This means the algorithm makes many assumptions that make it less likely to fit the true distribution of the data well and thus sacrifices some predictive power in exchange for accuracy estimates being less variable.

NB models are ‘black box’ models as one cannot extract a formula or sectioning point that can be easily applied without software (Seedat, van Niekerk, et al., 2009b). They would thus need to be far more accurate than other model types to justify their use.

When compared to discriminant models, sample size analyses showed that minimum mean error for NB models was higher, but error variance was lower. This was expected given the high bias, low variance nature of NB classifiers. At a sample size close to the size of the training dataset, variance for the discriminant and NB models did not very much. As such, considering the training sample size, discriminant models were favoured over NB models.

Univariate NB models for AD, THD, BB, TPB and HHD had either the highest or joint highest *k*-fold CV accuracies and kappa values. Univariate NB models did not suffer from overfitting (training and *k*-fold kappa values for all models were very similar with no obvious decline in accuracy). However univariate models including TAD, MCL, LCL, TPB and GB all presented with large sex biases. Sometimes the male accuracy was much higher and in other cases the female accuracy was much higher, but the presence of any significant sex bias made a model undesirable, especially when compared to one which had less notable bias.

The only multivariate NB model which was more successful than discriminant models was ShoulderNb. The accuracy was only 0.4% higher than for the discriminant model but sex bias was 5.2% compared to only 1.0% for the discriminant model. Therefore, the discriminant model was selected over the NB model.

Although NB models often had the highest kappa values or low sex bias, they were always outdone by discriminant models which tended to combine high accuracies and lower sex biases. Sample size analysis also showed the NB algorithm to be especially useful when sample sizes were small. Despite its merits, NB was not the best algorithm to suit the dataset or intended application for this study.

6.3.5 KNN

The KNN algorithm is an instance-based learner that categorises data points based on their proximity to other instances in the training dataset (Guo et al., 2003). The algorithm is easily 'fooled' by irrelevant attributes that obscure more important features often leading to overfitting with high dimensional data (Cunningham & Delany, 2021). Converse to NB, KNN is a high variance, low bias classifier (like decision trees).

The high variance, low bias nature of KNN classifiers was seen during sample size analysis. At larger sample sizes, the KNN algorithm achieved very low error however the width of the confidence interval was relatively wide. A larger dataset than what was used in this study would be ideal to achieve minimal error and variance with a KNN classifier.

In most cases the trained KNN algorithm provided univariate sex prediction models with accuracies which were average relative to other algorithm types. Only TADknn and GBknn models generated the joint highest *k*-fold CV accuracies for their respective predictors. The TADknn model was not selected as the best fit TAD model as the TADdiscr model achieved the same accuracy but with lower sex bias. The GBknn model was not selected as best fit because the improvement in accuracy of 0.6% could not justify the use of a 'black box' model with high variance over a far more reliable and simplistic DFA model.

All multivariate KNN joint models were overfit. Their training accuracy was 100% whilst their cross-validation accuracies were much lower. These models are thus not considered as potential best fit models. The models likely overfit because they were 'fooled' by irrelevant predictors which caused to model to be unable to generalise well to new data.

The APminkowskiKNN model was the most accurate AP model. It did not overfit, had the highest k-fold CV accuracy and the sex bias was small. It was however not selected as the best fit AP model as the DFA model had even less bias, and its slightly lower accuracy could be justified given the relative model complexity and ease of use.

The KNN algorithm may be a useful tool for sex estimation but requires much larger sample sizes to be maximally effective and reliable as a sex classifier.

RECOMMENDATIONS

Differences in accuracy between models trained using different machine learning algorithms were mostly small and the increased predictive power of models trained using more complex algorithms was never sufficient to warrant their use over more simplistic algorithms. As such, the continued use of DFA models for metric sex estimation is endorsed.

Increasing training sample size generally causes an increase in prediction accuracy up until a plateau is reached when the maximal accuracy is achieved (Balki et al., 2019). Algorithms like decision trees, ensemble methods and KNN were shown to require larger datasets than what was available in this study to function maximally. Increasing the size of the training sample may improve the predictive power and reliability of accuracy estimates for these algorithm types and make them viable alternatives to DFA.

The maximal possible accuracy that any algorithm or model could achieve was limited by the nature of the predictor variables and the dataset. Male and female measurement values overlap considerably because sex exists on a continuum in the human skeleton (Mircea, 2016). This makes accurate classification of individuals within the region of overlap difficult as has been noted by several published sources (Asala, 2002; Barrier, 2007; Brzobohatá et al., 2016; Jerković et al., 2020). The use of multiple predictors can aid in improving separation of the classes or cause redundancy when too many variables are included (Iguyon & Elisseeff, 2003; Vabalas et al., 2019). All the predictors in this study shared the same positive linear relationship with sex thus making the addition of multiple variables less valuable than if some predictors were differently correlated. Future work to identify predictors which correlate differently with sex may aid in improving predictive power of models. An example might be carrying angle of the elbow which is smaller for males than females. Another strategy may be to standardize the data using log or other normalization techniques (Feng et al., 2014; Curran-Everett, 2018) or to transform the available data using division by a specific denominator or other scaling methods before algorithm training.

Many algorithms are considered to be 'black boxes' as they cannot be manually applied without the help of software and a computer. These algorithms can be made more accessible to users by designing an app like CADOES (Coelho & Curate, 2019) or FORDISC (Ousley & Jantz, 2013) which allows a user to upload their data and receive an estimate of sex without needing to do any computation themselves.

6.4 COMPARATIVE DISCUSSION

Note: All accuracies are displayed with 1 decimal point for the sake of consistency within this section as most comparative sources only quote accuracies to a single decimal point.

The premise of this study was to determine whether ancestry-independent postcranial osteometric sex estimation models using a demographically representative, pooled South

African reference sample could be as accurate or more accurate than ancestry-specific discriminant functions.

The estimation of sex as part of a biological profile is vital in the correct identification of human skeletal remains (Konigsberg, Algee-Hewitt & Steadman, 2009) and many ancestry-specific sex estimation equations exist which are accurate to use on the population group which they are designed for (Steyn & Işcan, 1997, 1999; Asala, Bidmos & Dayal, 2004; Patriquin, Steyn & Loth, 2005; Steyn & Patriquin, 2009; MacAluso, 2011). The problem is, all ancestry-specific sex estimation equations rely on the "fundamental assumption" that the inclusion of ancestry will significantly improve the classification precision of the equation (Albanese et al., 2016). The biggest drawback to ancestry-specific equations is that ancestry must be known before an equation can be selected and applied to the task of estimating sex. The determination of ancestry itself is often based on further assumptions (Işcan & Steyn, 2013; Gannett, 2014) which make the assignment of ancestry to unknown skeletal remains anything from uncertain to impossible depending on which skeletal elements are available (Liebenberg, L'Abbé & Stull, 2015; Liebenberg et al., 2019). For example, multivariate postcranial ancestry estimation using the scapula (46%), or humerus (57%) would introduce a large degree of error (Liebenberg, L'Abbé & Stull, 2015).

Including ancestry as a prerequisite to sex estimation seems to cause many problems; but would SA ancestry-independent methods work equally well on a Dutch, Japanese or Egyptian population? Some authors claim that the "degree of sexual dimorphism varies greatly between populations" and advise that population-specific equations should be employed (Bidmos & Dayal, 2004; Bidmos & Mazengenya, 2021). Some studies had relative success in generating globally applicable standards, whilst others provided evidence to support the continued use of population-specific alternatives (Macho, 1990; Bidmos & Dayal, 2004; Steyn & Patriquin, 2009; MacAluso, 2010; Dillon, 2014; Kotěrová et al., 2017).

Pre-existing ancestry-specific and ancestry-independent SA and international sex estimation equations for different skeletal elements have been compared to the newly derived equations.

6.4.1 PELVIC MODELS - TABLE 6-2

The pelvis is widely considered to be one of the most useful skeletal elements for sex estimation thanks to its role in parturition (Bidmos, Gibbon & Štrkalj, 2010). Many sex estimation methods focus on pubic bone morphology (Lovell, 1989; Sutherland & Suchey, 1991) but "non-canal pelvic regions" were shown to display little significant differences between males and females across various geographical regions when compared to the birth canal itself (Kurki & Decrausaz, 2016). Consequently, it makes sense to hypothesise that since the acetabulum does not form part of the birth canal, the factors which would traditionally be cited as reasons for dimorphism in the pelvis, like parturition, have very little baring on any differences seen in the acetabulum. Despite this, the acetabulum has been described as "one of the most diagnostic pelvic elements for sex assessment" by Bubalo et al. (2019) and was specifically selected for this study due to the relatively high survival rate of the sciatic region of the pelvis (Stojanowski, Seidemann & Doran, 2002; Bubalo et al., 2019).

The best fit TAD model, TADdiscr (Table 5-11), had an estimated sex classification accuracy of 84.4%. This was higher than estimates for highly sex biased SAEA and similar to, but slightly higher than, estimated accuracies for SAAA functions (Patriquin, Steyn & Loth, 2005; Steyn & Patriquin, 2009). A proposed global function provided by Steyn & Patriquin (2009) which was derived using a pooled sample of SAAA, SAEA and Greeks also failed to achieve a better accuracy estimate than the newly derived model.

TAD, as defined by Bubalo et al. (2019), had not yet been tested on a South African sample. TAD proved to have better discriminatory power than AD in a Croatian sample but when measured and used to derive models for the South African population, it was found to be less useful than AD (82.8% accuracy for TAD compared to 84.4% for AD). It was hoped that combining AD and TAD would garner a pelvic model with better accuracy than AD alone but AD+TAD was not significantly more accurate. This was likely because AD and TAD proved to have a strong positive linear correlation with each other which resulted in less separation of classes than had been hoped for.

There may still be an unexplored relationship between TAD and AD. Future work to build models with transformed data (for example: $\frac{TAD}{TAD+AD}$ or $\left(TAD^2 + \sqrt{AD}\right)$) may uncover a model with superior sex predicting capabilities.

Bubalo et al. (2019) reported higher accuracies for all acetabular models than this study. This was likely because the Croatian population, which was used to build a reference sample, was more homogenous and less ancestrally diverse than the South African population and because all individuals in the Croatian study died traumatically in the Croatian war of independence over a period of only 4 years compared to the more than 100-year period for individuals in this study (Bubalo et al., 2019). Secular trend was shown to cause an increase in overlap between male and female measurements in this study and therefore may have been influential in the accuracy disparity between this study and Bubalo et al. (2019).

In conclusion, the best fit AD, TAD, and Pelvic models derived in this study were more accurate and less biased towards one sex than previously published ancestry-specific South African methods. The removal of ancestry as a prerequisite factor did not reduce the accuracy of AD as a predictor of sex and as such this author sees no justification for continued use of ancestry-specific equations for the acetabulum.

Table 6-2 Comparison of reported accuracies from different sources for measurements of the pelvis.

	AD			**TAD**			**Pelvis (AD + TAD)**			
Group	Male	Female	All	Male	Female	All	Male	Female	All	**Source**
SA	83.9	84.8	84.4	83.5	82.1	82.8	83.9	85.2	84.6	This Study
Croatian	90.0	84.0	87.0	88.0	88.0	88.0	88.0	88.0	88.0	(Bubalo et al., 2019)
SAEA	77.0	86.0	81.5	-	-	-	-	-	-	(Patriquin, Steyn & Loth, 2005)
SAAA	88.9	78.0	83.5	-	-	-	-	-	-	
SAEA	77.2	85.9	81.6	-	-	-	-	-	-	(Steyn & Patriquin, 2009)
SAAA	88.9	78.0	83.5	-	-	-	-	-	-	
Pooled*	80.5	84.4	82.5	-	-	-	-	-	-	

Pooled* SAAA, SAEA and Greeks from Crete

6.4.2 FEMORAL MODELS - TABLE 6-3 & TABLE 6-4

It is generally recommended that one should use the skull for sex assessment when the pelvis is unavailable. However, Spradley & Jantz (2011) set out to disprove this theory. They aimed to produce an "objective hierarchy of sexing effectiveness" of skull and postcranial elements to "test the notion that the skull is better" (Spradley & Jantz, 2011). Joint size of long bones – specifically the femur, tibia, and humerus – proved to be most useful and outperformed the crania.

The femur is one of the most dimorphic skeletal elements thanks to its biomechanical relationship with the pelvis (Albanese, Eklics & Tuck, 2008; Christensen, Passalacqua & Bartelink, 2019). The proximal femur is particularly useful and the concentration of most femoral dimorphism due to its direct relationship and articulation with the pelvis (Curate et al., 2016). Another contributor to the differences between male and female pelvic bones is dimorphic muscle attachments relevant in body weight transmission (Kim, Kwak & Han, 2013).

As expected, the proximal femur outperformed distal femoral mensurations in this study (Table 5-11). Accuracies for models using proximal femoral mensurations were more than 3% higher than models utilizing the distal femur. The 2 proximal femoral mensurations, FHD and THD, proved to be equally accurate for sex estimation and sectioning points for FHD and THD models were almost identical. This is a useful finding as whilst FHD is widely studied, THD is less often assessed (Table 6-3). THD models can therefore be used as a substitute to FHD models when the femoral head is damaged in a manner that prevents the accurate measurement of FHD.

Steyn & Işcan (1997) reported sex prediction accuracies of 88.9% for SAEA using FHD. This was ~2% higher than what this study achieved using a pooled South African sample. It is important to note that the reference sample in Steyn & Işcan (1997) consisted of only 106 individuals and no cross-validation techniques were applied. Direct resubstitution of training data often causes inflation of accuracy estimates (Vabalas et al., 2019). Another study reported an accuracy of 82.6% for FHD using a reference sample of 220 SAAA individuals

(Asala, Bidmos & Dayal, 2004). Given the sample size of 220, the width of the confidence interval would theoretically be quite wide at around 1.8% (Figure 5-17). The new ancestry-independent FHDdiscr model was more than 4% more accurate than the SAAA-specific model and was therefore superior.

SAAA- and SAEA-specific equations need to be cross-validated on larger sample sizes and then compared to ancestry-independent results to conclusively determine which models perform best on the South African population. Acknowledging the limitations of the comparative studies, we can tentatively observe that the new ancestry-independent discriminant function for FHD is an accurate and useful alternative to the ancestry-dependent functions.

New FHD and THD estimated model accuracies were compared to international sources. FHDdiscr and THDdiscr produced higher accuracy estimates than discriminant functions derived for a Portuguese reference sample of 200 individuals (Curate, Umbelino, et al., 2017). Spradley & Jantz (2011) also studied FHD and reported a discriminant function sexing accuracy of 86.0% and a sectioning point of 44mm using an American reference sample. This was very similar to the accuracy of 86.9% for FHDdiscr although the sectioning point for FHDdiscr was not the same (43.3mm - Table 5-11). It was unclear why accuracies from the Portuguese study were so much lower (potentially due to the wide skeletal age range of 20-89 years), but this does confirm that the new models derived in this study are in line with international accuracies, despite South Africa's highly heterogenous population.

One South African study investigated sex estimation using a pooled SA sample. Krüger, L'Abbé & Stull (2017) used DFA to derive ancestry-independent sex estimation equations for a South African reference sample of 360 individuals which incorporated an equal number of SAAA, SAEA and SAMA individuals compared to the demographically representative sample used in this study (Krüger, L'Abbé & Stull, 2017). The reported accuracy for their FHD discriminant function was almost the same (0.9% lower) as the estimated accuracy of FHDdiscr. This shows that, in the case of FHD, the exact demography of the reference sample did not impact the final model accuracy. Unfortunately, sectioning points are not reported in Krüger, L'Abbé & Stull (2017) so a comparison cannot be made.

Table 6-3 Comparison of reported accuracies from different sources for measurements of the proximal femur.

	FHD			THD			
Group	Male	Female	All	Male	Female	All	**Source**
SA	86.2	87.6	86.9	86.3	87.2	86.7	New
Portuguese	76.0	86.0	81.0	83.0	71.0	77.0	(Curate, Umbelino, et al., 2017)
American	-	-	86.0	-	-	-	(Spradley & Jantz, 2011)
SAEA	87.5	84.0	88.9	-	-	-	(Steyn & Işcan, 1997)
SAAA	-	-	82.6	-	-	-	(Asala, Bidmos & Dayal, 2004)
SA (equal)	87.0	86.0	86.0	-	-	-	(Krüger, L'Abbé & Stull, 2017)

Sex estimation accuracies for distal femoral models were inferior to those for proximal femoral models. MCLdiscr and LCLdiscr were the worst performing univariate models in this study. Reported accuracies from comparative local and international sources in Table 6-4 were inconsistent. Some models seemed to perform similarly to the new models whilst others like Curate, Umbelino, et al. (2017), with 63.0% accuracy for BB or Steyn & Işcan (1997), with 90.5% for BB were either by far inferior or by far superior. These inconsistent findings make comparisons to the new model difficult.

The multivariate FemurDiscr model used a combination of 5 predictors which have not previously been studied together in South Africa. The 5-predictor FemurDiscr model had an estimated model accuracy of 87.5% (Table 5-11) and was only slightly more accurate than univariate FHDdiscr (86.75%). Other SA studies have reported similar accuracies to those estimated for FemurDiscr. The best femoral model reported by Krüger et al. (2017) was a stepwise LDA femoral model which attained 86.0% accuracy whilst the best femoral model reported in Asala et al. (2004) was a 5-predictor linear discriminant model which achieved 85.1% accuracy and Steyn & Işcan (1999) reported an accuracy of 88.6% for their 3-predicotor femoral model.

In isolation from other sources, the newly derived univariate models for the distal femur were the least accurate of all models in this study and would thus be recommended only as a last resort if none of the other skeletal elements examined in this study were available for sex estimation. Univariate models for the proximal femur were much more accurate and are thus preferable over both pelvic models and distal femoral models. Whilst the 5-predictor femoral model was more accurate than its univariate compatriots, TPBdiscr, which will be discussed next, was able to achieve almost the same accuracy with only 1 predictor.

Table 6-4 Comparison of reported accuracies from different sources for measurements of the distal femur.

	MCL			BB			LCL			
Group	Male	Female	All	Male	Female	All	Male	Female	All	**Source**
All	79.1	81.5	80.3	83.1	83.7	83.4	78.7	82.3	80.5	New
Portuguese	-	-	-	71.0	55.0	63.0	84.0	91.0	87.5	(Curate, Umbelino, et al., 2017)
Korean	-	-	82.7	-	-	-	-	-	81.7	(Kim, Kwak & Han, 2013)
SAEA	-	-	-	89.3	91.8	90.5	-	-	-	(Steyn & Işcan, 1997)
SAAA	-	-	80.5	-	-	81.5	-	-	75.6	(Asala, Bidmos & Dayal, 2004)
Pooled	-	-	-	-	-	81.0	-	-	-	(Krüger, L'Abbé & Stull, 2017)

6.4.3 TIBIA MODELS - TABLE 6-5

The lower limb, including the tibia, responds to hormones and environmental stressors such as physical load which contribute to sexual dimorphism (Carlson & Marchi, 2014; Brzobohatá et al., 2016). TPB, the only tibial mensuration in this study, proved to be one of the best univariate predictors of sex, garnering an estimated accuracy of 87.2%. This accuracy was comparable with reported accuracies for a SAEA (88.4%), pooled South African (86.0%), Croatian (87.7%) and American (88.0%) sample explored in Table 6-5. Given the variability of accuracy estimates at smaller sample sizes (Figure 5-17), the variance in accuracy between these local and international studies was likely negligible.

No comparable discriminant functions were derived for TBP using a SAAA or SAMA reference sample, but the accuracy achieved by the new ancestry-independent model was similar to (although slightly lower than) the accuracy for the SAEA model (Steyn & Işcan, 1997). As previously noted, Steyn and Işcan (1997) used a small sample of only 106 individuals so a difference of 1.2% is within the bounds of expected variance.

The new ancestry-independent TPB model was a useful and accurate univariate predictor of sex and is recommended in place of the existing ancestry-dependent model, especially considering the larger reference sample used to derive the new model and reduced sex bias.

Table 6-5 Comparison of reported accuracies from different sources for TPB.

	TBP			
Group	Male	Female	All	**Source**
All	86.6	87.7	87.2	This study
SAEA	86.8	90.9	88.4	(Steyn & Işcan, 1997)
All	84.0	88.0	86.0	(Krüger, L'Abbé & Stull, 2017)
Croatian	89.0	85.9	87.8	(Šlaus et al., 2013)
American	-	-	88.0	(Spradley & Jantz, 2011)

6.4.4 SCAPULA MODELS - TABLE 6-6

GL and GB were identified as the most dimorphic of all skeletal measurements investigated in this study and as such, it was hypothesised that these would be the best predictors of sex. Evidence has been presented to support the thesis that the glenoid cavity is preserved at a higher rate than most other postcranial traits with the "possible exception of long bone shaft fragments" (Stojanowski, Seidemann & Doran, 2002). Exceptional preservation and sexual dimorphism were strong motivators for the use of the glenoid fossa in sex estimation.

As expected, given that it was the most dimorphic mensuration (Table 5-6), GL produced the univariate sex estimation model with the highest accuracy. GB was slightly less successful than anticipated but when combined, GL and GB produced the most accurate multivariate model aside from the AP and Optimal models.

GL and GB have only been explored in a SAAA reference sample of 120 individuals within South Africa by MacAluso (2011). MacAluso (2011) utilized logistic regression to derive their sectioning points, whilst the current study used linear DFA for both mensurations. Average measurement values from MacAluso (2011) were very similar to those reported in this study and sectioning points for GB were almost identical (25.16mm vs 25.18mm). However, the sectioning point for GL was considerably different.

The MacAluso (2011) sectioning point for GL can be derived as 31.07mm whilst the sectioning point in this study was considerably higher at 36.76mm (Table 5-11). The new GBdiscr model was slightly less accurate than the MacAluso (2011) ancestry-dependent model whilst GLdiscr was 2.1% more accurate. The differences in accuracy between new models and existing models may largely be attributed to the large variance in error which is expected at a sample size of 120 (Figure 5-17) and the big difference in sectioning point values for GL models.

The multivariate ScapulaDiscr (90.3%) model outperformed its respective best univariate component, GLdiscr (88.7%) and was more accurate than a similar model presented by MacAluso (2011) which achieved an accuracy of 88.3% using the area of the glenoid fossa.

It has been established that the glenoid fossa is a useful skeletal element for sex estimation and GLdiscr has proven to be the best univariate predictor of sex. Since the new ancestry-independent models were trained using a much more reliable, larger dataset, and achieved similar and better estimation accuracies, they are recommended in place of the existing ancestry-specific models.

Table 6-6 Comparison of reported accuracies from different sources for scapular measurements.

	GL			GB			
Group	Male	Female	All	Male	Female	All	**Source**
All	88.1	89.2	88.7	86.5	82.8	84.6	This study
SAAA	86.7	86.7	86.7	83.3	88.3	85.8	(MacAluso, 2011)

6.4.5 HUMERUS MODELS - TABLE 6-7

Alongside GB and TPB, EB was one of the most accurate univariate predictors of sex in this study. Reported accuracies for SAAA and SAEA from Steyn and Işcan (1999) for EB and HHD models were higher than accuracies achieved in the current study. The difference in accuracy can likely be attributed to sample size. The sample sizes of 88 for SAAA and 104 for SAEA in Steyn and Işcan (1999) mean that the reference sample was most likely too small to eliminate the high error variance associated with small sample sizes and the results were likely highly sample-dependent (Vabalas et al., 2019). Both EB and HHD models for SAEA were considerably sex biased with reported accuracies more than 10% higher for females than males (Table 6-7).

Krüger et al. (2017), as previously mentioned, studied a pooled South African reference sample. Accuracies for the new EB and HHD models were higher than accuracies reported by Krüger et al. (2017) for the same predictor variables.

The new multivariate HumerusDiscr, combining EB and HHD, was trained and achieved a sex estimation accuracy of 89.98%. This was superior to accuracies for either of the model's univariate components. When compared to equations using the same predictors in Steyn & Işcan (1999), which achieved an accuracy of 92.2%, the new model seems to be inferior. However, a different study validated the accuracies reported by Steyn & Işcan (1999) using reference samples from UCT, Dart and Pretoria skeletal repositories and results varied from 88.5-95.5% with an overall average of 90.5% (Robinson & Bidmos, 2009). Most of the current study sample consisted of individuals from the Dart collection which had the lowest validation result at 88.5%. Given the vast variability in accuracies depending on the sample used to test the same function, it is uncertain whether the new function was more accurate but there is little justification to favour the ancestry-dependent function especially given the added error associated with predetermination of ancestry.

Table 6-7 Comparison of reported accuracies from different sources for humerus measurements.

	EB			HHD			
Group	Male	Female	All	Male	Female	All	**Source**
SA	87.3	87.3	87.3	82.7	84.1	83.4	This study
SAAA	86.0	91.1	88.6	93.0	88.9	90.9	(Steyn & Işcan, 1999)
SAEA	83.6	95.8	89.7	78.2	89.8	83.7	
SA (equal)	-	-	85.0	-	-	78.0	(Krüger, L'Abbé & Stull, 2017)
American	-	-	86.0	-	-	86.0	(Spradley & Jantz, 2011)

SUMMARY OF FINDINGS

Sample size is an important consideration when training machine learning algorithms to estimate sex, as an insufficient reference sample will not be an accurate representation of the population and will result in a model that does not generalize well to new data (Brain & Webb, 1999). The sample size for this study was shown in Figure 5-17 to be sufficient, whereas the sample sizes employed by most SA studies (Steyn & Işcan, 1997, 1999; Asala, Bidmos &

Dayal, 2004; MacAluso, 2011), with the exception of Patriquin (2005), were not large enough to train accurate and reliable discriminant function models with minimal error variance.

Most of the new models presented notably less sex bias than existing SA and international models. This was considered to be a beneficial trait as sacrificing the accuracy for one group in order to maximise the accuracy of another group leads to classification inequality.

As previously mentioned, the principal concern with ancestry-specific equations is the need to assign the unknown skeletal remains to an ancestry group prior to sex estimation. The new ancestry-independent functions remove this source of potential error by eliminating the necessity to allocate unknown skeletal remains to a prescribed ancestry group. All ancestry-specific equations rely on the cardinal postulation that the inclusion of ancestry improves the precision of sex estimation, so the removal of group-specificity was expected to reduce prediction accuracy. Contrary to this assumption, within the expected CI for accuracy estimated at given sample sizes, the new ancestry-independent functions were often equivalently accurate or more accurate than the existing ancestry-dependent equations. Considering the added benefit of abolishing group-specificity, the new equations are the expedient option.

Most available methods were ancestry-specific with the exception of Krüger et al. (2017). Krüger et al. (2017) used a pooled South African sample consisting of equally sized cohorts from the 3 major SA ancestry groups whereas this study was conducted with a demographically representative sample consisting of mainly SAAA individuals. Accuracies for the new models were higher or equal to accuracies reported by Krüger et al. (2017) but most likely favour SAAA, the majority group. If the ancestry of an unknown individual is suspected to be SAEA or SAMA, given the division of ancestry in Krüger et al. (2017)'s pooled sample, their models would potentially perform better as the cohort of SAEA and SAMA individuals in their study was slightly larger (although their overall sample size of 350 was smaller).

The new ancestry-independent univariate sex prediction models are all good alternatives to ancestry-specific models. The most accurate and thus most recommended, univariate models are GL, TPB and EB. These and other univariate models are useful for quick analysis and are especially expedient when limited or fragmentary skeletal material is available for demographic analyses (Kelley, 1979; Asala, Bidmos & Dayal, 2004).

When entire undamaged bones or joints are available, multivariate equations can increase the accuracy and reliability of sex estimation (Krüger, L'Abbé & Stull, 2017). The least successful multivariate models were the lower limb models for the Pelvis, Hip and Knee. None of these models had notably higher prediction accuracies than their most successful univariate component. The best multivariate model, if all requisite elements are available, was the Optimal model with an estimated accuracy of 92.7%. This model combines the most useful predictors to create a highly accurate, minimally redundant sex classifier. Alternate highly accurate multivariate models are ScapulaDiscr and HumerusDiscr both of which markedly outperform ShoulderDiscr and their univariate components. The 'Optimal' model uses a new combination of predictors which has not yet been explored in SA, so it is not possible to make comparisons with other sources in the literature.

CHAPTER 7: CONCLUSION

12 osteometric parameters were measured from the cadaveric skeletal remains of 650 South Africans. All parameters proved to be sexually dimorphic with male measurement values being on average larger and more variable than female values. A large region of overlap between male and female values was identified and age-correlated osteological changes were shown to increase overlap between older females and small males. All sex classification models struggled to accurately predict the sex of individuals which fell into the region of overlap. The inclusion of predictor variables which do not have a positive linear correlation with sex or, the transformation of existing predictors before ML algorithm training may help to improve separation between the sexes and increase the predictive power of ML models.

Different ML techniques were used to develop univariate and multivariate sex classification models. A *post hoc* sample size analysis concluded that a sample size of 400 or more individuals was sufficient to yield accurate and reliable sex prediction models and accuracies for DFA and NB algorithms but not for decision tree, ensemble or KNN algorithms. The algorithm used had very little impact on the accuracy of the final trained model. DFA models were chosen as best fit for all univariate and multivariate predictor combinations due to their consistently high accuracies, low sex biases and relative model simplicity. Univariate DFA models achieved sex prediction accuracies of 80.5-90.0%. The most accurate univariate predictors were GLdiscr (88.7%), EBdiscr (87.3%) and TPBdiscr (87.2%). Multivariate models scored accuracies between 84.6% and 92.8%. The most accurate multivariate model was the 12-predictor AP model with 92.8% accuracy but the less complex 3-predictor Optimal model, combining LCL, TPB and GL, achieved 92.7%. The less complex Optimal model was favoured as it was able to combine less coefficients and variables to achieve an almost identical result.

When compared to existing sex prediction methods, the new ancestry-independent discriminant models were recommended in place of ancestry-dependent models as they achieved similar and better prediction accuracies. Additionally, the new models were trained with a larger and thus more reliable dataset than existing methods.

REFERENCES

Adhikari, M. 2006. 'God made the white man, god made the black man…': Popular racial stereotyping of coloured people in apartheid South Africa. *South African Historical Journal*. 55(1):142–164. DOI: 10.1080/02582470609464935.

Aguinis, H., Gottfredson, R.K. & Joo, H. 2013. Best-Practice Recommendations for Defining, Identifying, and Handling Outliers. *Organizational Research Methods*. 16(2):270–301. DOI: 10.1177/1094428112470848.

Akachi, Y. & Canning, D. 2007. The height of women in Sub-Saharan Africa: The role of health, nutrition, and income in childhood. *Annals of Human Biology*. 34(4):397–410. DOI: 10.1080/03014460701452868.

Albanese, J., Eklics, G. & Tuck, A. 2008. A metric method for sex determination using the proximal femur and fragmentary hipbone. *Journal of Forensic Sciences*. 53(6):1283–1288. DOI: 10.1111/j.1556-4029.2008.00855.x.

Albanese, J., Tuck, A., Gomes, J. & Cardoso, H.F.V. 2016. An alternative approach for estimating stature from long bones that is not population- or group-specific. *Forensic Science International*. 259:59–68. DOI: 10.1016/j.forsciint.2015.12.011.

Alblas, A., Greyling, L.M. & Geldenhuys, E.-M. 2018. Composition of the Kirsten Skeletal Collection at Stellenbosch University. *South African Journal of Science*. 114(1/2):1–6. DOI: 10.17159/sajs.2018/20170198.

Ali, A., Ahmed, T., Ayub, A., Dano, S., Khalid, M., El-Dassouki, N., Orchanian-Cheff, A., Alibhai, S., et al. 2020. DOI: 10.1111/ctr.13832.

Alloghani, M., Al-Jumeily, D., Mustafina, J., Hussain, A. & Aljaaf, A.J. 2020. *Supervised and Unsupervised Learning for Data Science*. M.W. Berry, A. Mohamed, & B.W. Yap, Eds. (Unsupervised and Semi-Supervised Learning). Cham: Springer International Publishing. DOI: 10.1007/978-3-030-22475-2.

Arendse, L.M. 2018. Stature Estimation: Evaluating Regression Formulae for Different Population Groups in South Africa. University of Cape Town.

Arnastauskaitė, J., Ruzgas, T. & Bražėnas, M. 2021. An exhaustive power comparison of normality tests. *Mathematics*. 9(7):788. DOI: 10.3390/math9070788.

Asala, S.A. 2001. Sex determination from the head of the femur of South African whites and blacks. *Forensic Science International*. 117(1–2):15–22. DOI: 10.1016/S0379-0738(00)00444-8.

Asala, S.A. 2002. The efficiency of the demarking point of the femoral head as a sex determining parameter. *Forensic Science International*. 127(1–2):114–118. DOI: 10.1016/S0379-0738(02)00114-7.

Asala, S.A., Bidmos, M.A. & Dayal, M.R. 2004. Discriminant function sexing of fragmentary femur of South African blacks. *Forensic Science International*. 145(1):25–29. DOI: 10.1016/j.forsciint.2004.03.010.

Baker, M. & Penny, D. 2016. Is there a reproducibility crisis? *Nature*. 533(7604):452–454. DOI: 10.1038/533452A.

Balci, Y., Yavuz, M.F. & Cağdir, S. 2005. Predictive accuracy of sexing the mandible by ramus flexure. *HOMO- Journal of Comparative Human Biology*. 55(3):229–237. DOI: 10.1016/j.jchb.2004.07.006.

Balki, I., Amirabadi, A., Levman, J., Martel, A.L., Emersic, Z., Meden, B., Garcia-Pedrero, A., Ramirez, S.C., et al. 2019. DOI: 10.1016/j.carj.2019.06.002.

Barrier, I.L.O. 2007. Sex determination from the bones of the forearm in a modern South African sample. University of Pretoria. Available: https://repository.up.ac.za/handle/2263/27036 [2021, October 30].

Baum, E.B. & Haussler, D. 1989. What Size Net Gives Valid Generalization? *Neural Computation*. 1(1):151–160. DOI: 10.1162/neco.1989.1.1.151.

Beam, A.L., Manrai, A.K. & Ghassemi, M. 2020. Challenges to the Reproducibility of Machine Learning Models in Health Care. *JAMA - Journal of the American Medical Association*. 323(4):305–306. DOI: 10.1001/jama.2019.20866.

Benecke, M. 2015. Six Forensic Entomology Cases: Description and Commentary. *Journal of Forensic Sciences*. 43(4):14309J. DOI: 10.1520/jfs14309j.

Berrar, D. 2018. Bayes' Theorem and Naïve Bayes Classifier. In *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*. V. 1–3. S. Ranganathan, M. Gribskov, K. Nakai, & C. Schöenbach, Eds. Catanzro, Italy: Elsevier. 403–412. DOI: 10.1016/B978-0-12-809633-8.20473-1.

Bidmos, M.A. & Dayal, M.R. 2004. Further Evidence to Show Population Specificity of Discriminant Function Equations for Sex Determination Using the Talus of South African Blacks. *Journal of Forensic Sciences*. 49(6):1–6. DOI: 10.1520/jfs2003431.

Bidmos, M.A. & Mazengenya, P. 2021. Accuracies of discriminant function equations for sex estimation using long bones of upper extremities. *International Journal of Legal Medicine*. 135(3):1095–1102. DOI: 10.1007/s00414-020-02458-y.

Bidmos, M.A., Steinberg, N. & Kuykendall, K.L. 2005. Patella measurements of South African whites as sex assessors. *HOMO- Journal of Comparative Human Biology*. 56(1):69–74. DOI: 10.1016/j.jchb.2004.10.002.

Bidmos, M.A., Gibbon, V.E. & Štrkalj, G. 2010. Recent advances in sex identification of human skeletal remains in South Africa. *South African Journal of Science*. 106(11–12):1–6. DOI: 10.4102/sajs.v106i11/12.238.

Bogin, B. 1995. *Patterns of Human Growth*. Third ed. Cambridge UK: Cambridge University Press. DOI: 10.1126/science.7716552.

Bowman, A.W. & Azzalini, A. 1999. Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-PLUS Illustrations. *Journal of the American Statistical Association*. 94(447):982. DOI: 10.2307/2670015.

Brain, D. & Webb, G.I. 1999. *On The Effect of Data Set Size on Bias And Variance in Classification Learning*. Sydney.

Breiman, L., Friedman, J.H., Olshen, R.A. & Stone, C.J. 1984. *Classification and regression trees*. 1st ed. Boca Raton: Routledge. DOI: 10.1201/9781315139470.

Brzobohatá, H., Krajíèek, V., Horák, Z. & Velemínská, J. 2016. Sexual dimorphism of the human tibia through time: Insights into shape variation using a surface-based approach. *PLoS ONE*. 11(11):e0166461. DOI: 10.1371/journal.pone.0166461.

Bubalo, P., Baković, M., Tkalčić, M., Petrovečki, V. & Mayer, D. 2019. Acetabular osteometric standards for sex estimation in contemporary Croatian population. *Croatian Medical Journal*. 60(3):221–226. DOI: 10.3325/cmj.2019.60.221.

Buikstra, J.E. & Uberlaker, D.H. 1994. *Standards: For data collection from human skeletal remains*.

Burr, D.B. 1997. Muscle strength, bone mass, and age-related bone loss. *Journal of bone and mineral research: the official journal of the American Society for Bone and Mineral Research*. 12(10):1547–1551. DOI: 10.1359/jbmr.1997.12.10.1547.

Büyüköztürk, Ş. & Çokluk-Bökeoğlu, Ö. 2008. Discriminant function analysis: Concept and application. *Egitim Arastirmalari - Eurasian Journal of Educational Research*. (33):73–92.

Bytheway, J.A. & Ross, A.H. 2010. A geometric morphometric approach to sex determination of the human adult os coxa. *Journal of Forensic Sciences*. 55(4):859–864. DOI: 10.1111/j.1556-4029.2010.01374.x.

Carlson, K.J. & Marchi, D. 2014. Introduction: Towards refining the concept of mobility. In *Reconstructing Mobility: Environmental, Behavioral, and Morphological Determinants*. V. 9781489974. 1–11. DOI: 10.1007/978-1-4899-7460-0_1.

Carvallo, D. & Retamal, R. 2020. Sex estimation using the proximal end of the femur on a modern Chilean sample. *Forensic Science International: Reports*. 2(100077). DOI: 10.1016/j.fsir.2020.100077.

de Castro, F. 2021. *fitmethis*. MATLAB Central File Exchange. Available: https://www.mathworks.com/matlabcentral/fileexchange/40167-fitmethis), MATLA [2021, July 29].

Christensen, A.M. & Crowder, C.M. 2009. Evidentiary standards for forensic anthropology. *Journal of Forensic Sciences*. 54(6):1211–1216. DOI: 10.1111/j.1556-4029.2009.01176.x.

Christensen, A.M., Passalacqua, N. V. & Bartelink, E.J. 2019. *Forensic Anthropology: Current Methods and Practice*. Second ed. E.A. Brown & L. Lima, Eds. London: Academic Press. DOI: 10.1016/b978-0-12-815734-3.00014-2.

Coelho, J.D. & Curate, F. 2019. CADOES: An interactive machine-learning approach for sex estimation with the pelvis. *Forensic Science International*. 302(109837). DOI: 10.1016/j.forsciint.2019.109873.

Cole, T.J. 2000. Secular trends in growth. In *Proceedings of the Nutrition Society*. V. 59. 317–324. DOI: 10.1017/S0029665100000355.

Compston, J.E., Vedi, S., Kaptoge, S. & Seeman, E. 2007. Bone remodelling rate and remodelling balance are not co-regulated in adulthood: Implications for the use of activation frequency as an index of remodelling rate. *Journal of Bone and Mineral Research*. 22(7):1031–1036. DOI: 10.1359/jbmr.070407.

Cornwall, J., Callahan, D. & Wee, R. 2016. Ethical Issues Surrounding the Use of Images From Donated Cadavers in the Anatomical Sciences. *Clinical Anatomy*. (29):30–36. DOI: 10.1002/ca.22644.

Cowgill, L.W., Eleazer, C.D., Auerbach, B.M., Temple, D.H. & Okazaki, K. 2012. Developmental variation in ecogeographic body proportions. *American Journal of Physical Anthropology*. 148(4):557–570. DOI: 10.1002/ajpa.22072.

Cunha, E. & Ubelaker, D.H. 2020. Evaluation of ancestry from human skeletal remains: a concise review. *Forensic Sciences Research*. 5(2):89–97. DOI: 10.1080/20961790.2019.1697060.

Cunningham, P. & Delany, S.J. 2021. K-Nearest Neighbour Classifiers 2nd Edition (with Python examples). *ACM Computing Surveys*. 54(6):1–6. DOI: 10.1145/3459665.

Curate, F., Coelho, J., Gonçalves, D., Coelho, C., Ferreira, M.T., Navega, D. & Cunha, E. (in press). A method for sex estimation using the proximal femur. *Forensic Science International*. 266:579.e1-579.e7. DOI: 10.1016/j.forsciint.2016.06.011.

Curate, F., Umbelino, C., Perinha, A., Nogueira, C., Silva, A.M. & Cunha, E. 2017. Sex determination from the femur in Portuguese populations with classical and machine-learning classifiers. *Journal of Forensic and Legal Medicine*. 52:75–81. DOI: 10.1016/j.jflm.2017.08.011.

Curate, F., Albuquerque, A., Ferreira, I. & Cunha, E. 2017. Sex estimation with the total area of the proximal femur: A densitometric approach. *Forensic Science International*. 275:110–116. DOI: 10.1016/j.forsciint.2017.02.035.

Curran-Everett, D. 2018. Explorations in statistics: The log transformation. *Advances in Physiology Education*. 42(2):343–347. DOI: 10.1152/ADVAN.00018.2018/SUPPL_FILE/TABLE_1_DATA.CSV.

Dabbs, G.R. & Moore-Jansen, P.H. 2010. A method for estimating sex using metric analysis of the scapula. *Journal of Forensic Sciences*. 55(1):149–152. DOI: 10.1111/j.1556-4029.2009.01232.x.

Dayal, M.R., Spocter, M.A. & Bidmos, M.A. 2008. An assessment of sex using the skull of black South Africans by discriminant function analysis. *HOMO- Journal of Comparative Human Biology*. 59(3):209–221. DOI: 10.1016/j.jchb.2007.01.001.

Dayal, M.R., Kegley, A.D.T., Goran, S., Bidmos, M.A. & Kuykendall, K.L. 2009. The History and Composition of the Raymond A . Dart Collection of Human Skeletons at the

University of the Witwatersrand, Johannesburg , South Africa. *American Journal of Physical Anthropology*. (140):324–335. DOI: 10.1002/ajpa.21072.

Deaton, A. 2007. Height, health, and development. *Proceedings of the National Academy of Sciences*. 104(33):13232–13237. DOI: 10.1073/pnas.0611500104.

Degusta, D. 1999. Fijian cannibalism: Osteological evidence from Navatu. *American Journal of Physical Anthropology*. 110(2):215–241. DOI: 10.1002/(SICI)1096-8644(199910)110:2<215::AID-AJPA7>3.0.CO;2-D.

Dent, B.B., Forbes, S.L. & Stuart, B.H. 2004. Review of human decomposition processes in soil. *Environmental Geology*. 45(4):576–585. DOI: 10.1007/s00254-003-0913-z.

Dietterich, T.G. & Kong, E.B. 1995. *Machine Learning Bias, Statistical Bias, and Statistical Variance of Decision Tree Algorithms*. Corvallis.

Dillon, A. 2014. Cranial sexual dimorphism and the population specificity of anthropological standards. The University of Western Australia. Available: https://api.research-repository.uwa.edu.au/files/3367364/Dillon_Alexandra_2014.pdf [2020, November 02].

Donaldson, A.E. & Lamont, I.L. 2014. Estimation of post-mortem interval using biochemical markers. *Australian Journal of Forensic Sciences*. 46(1):8–26. DOI: 10.1080/00450618.2013.784356.

Evert, L. & Rossouw, S.H. 2011. Unidentified Bodies in Forensic Pathology in South Africa.

Faerman, M., Nebel, A., Filon, D., Thomas, M.G., Bradman, N., Ragsdale, B.D., Schultz, M. & Oppenheim, A. 2000. From a dry bone to a genetic portrait: A case study of sickle cell anemiaThe contribution of the first and the second authors is equal. *American Journal of Physical Anthropology*. 111(2):153. DOI: 10.1002/(sici)1096-8644(200002)111:2<153::aid-ajpa2>3.3.co;2-f.

Feng, C., Wang, H., Lu, N., Chen, T., He, H., Lu, Y. & Tu, X.M. 2014. Log-transformation and its implications for data analysis. *Shanghai Archives of Psychiatry*. 26(2):105. DOI: 10.3969/J.ISSN.1002-0829.2014.02.009.

Fisher, R.A. 1936. THE USE OF MULTIPLE MEASUREMENTS IN TAXONOMIC PROBLEMS. *Annals of Eugenics*. 7(2):179–188. DOI: 10.1111/j.1469-1809.1936.tb02137.x.

Franklin, D. 2010. Forensic age estimation in human skeletal remains: Current concepts and future directions. *Legal Medicine*. (12):1–7. DOI: 10.1016/j.legalmed.2009.09.001.

Gangata, Hope. 2015. A proposed worldwide classification system for ways of sourcing of anatomical cadavers that is progressive towards the use of donated anatomical cadavers. *Edorium Journal of Anatomy and Embryology*. 2:20–25. DOI: 10.5348/A04-2015-6-ED-5.

Gangata, H., Ntaba, P., Akol, P. & Louw, G. 2010. The reliance on unclaimed cadavers for anatomical teaching by medical schools in Africa. *Anatomical Sciences Education*. 3(4):174–183. DOI: 10.1002/ase.157.

Gannett, L. 2014. Biogeographical ancestry and race. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*. 47(PA):173–184. DOI: 10.1016/j.shpsc.2014.05.017.

German, A., Livshits, G., Peter, I., Malkin, I., Dubnov, J., Akons, H., Shmoish, M. & Hochberg, Z. 2015. Environmental rather than genetic factors determine the variation in the age of the infancy to childhood transition: A twin study. *Journal of Pediatrics*. 166(3):731–735. DOI: 10.1016/j.jpeds.2014.11.047.

Gibbon, V., Paximadis, M., Štrkalj, G., Ruff, P. & Penny, C. 2009. Novel methods of molecular sex identification from skeletal tissue using the amelogenin gene. *Forensic Science International: Genetics*. 3(2):74–79. DOI: 10.1016/j.fsigen.2008.10.007.

Gibbon, V.E., Penny, C.B., Štrkalj, G. & Ruff, P. 2009. Minimally invasive bone sampling method for DNA analysis. *American Journal of Physical Anthropology*. 139(4):596–599. DOI: 10.1002/ajpa.21048.

Golland, P., Grimson, W.E.L., Shenton, M.E. & Kikinis, R. 2000. Small sample size learning for shape analysis of anatomical structures. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. V. 1935. 72–82. DOI: 10.1007/978-3-540-40899-4_8.

González, S., García, S., del Ser, J., Rokach, L. & Herrera, F. 2020. A practical tutorial on bagging and boosting based ensembles for machine learning: Algorithms, software tools, performance study, practical perspectives, and opportunities. *Information Fusion*. 64(July):205–237. DOI: 10.1016/j.inffus.2020.07.007.

Grivas, C.R. & Komar, D.A. 2008. Kumho, Daubert, and the Nature of Scientific Inquiry: Implications for Forensic Anthropology*. *Journal of Forensic Sciences*. 53(4):771–776. DOI: 10.1111/J.1556-4029.2008.00771.X.

Guo, G., Wang, H., Bell, D., Bi, Y. & Greer, K. 2003. KNN model-based approach in classification. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. 2888:986–996. DOI: 10.1007/978-3-540-39964-3_62.

Gürses, İ.A., Ertaş, A., Gürtekin, B., Coşkun, O., Üzel, M., Gayretli, Ö. & Demirci, M.S. 2019. Profile and Motivations of Registered Whole-Body Donors in Turkey: Istanbul University Experience. *Anatomical Sciences Education*. 12(4):370–385. DOI: 10.1002/ase.1849.

Haglund, W., Connor, M. & Scott, D. 2001. The Archaeology of Contemporary Mass Graves. *Historical Arcahaeology*. 35(1):57–69. DOI: 10.1179/0032044714z.00000000020.

Hawkins, D.M. 2004. The Problem of Overfitting. *Journal of Chemical Information and Computer Sciences*. 44(1):1–12. DOI: 10.1021/ci0342472.

Hawley, N.L., Rousham, E.K., Norris, S.A., Pettifor, J.M. & Cameron, N. 2009. Secular trends in skeletal maturity in South Africa: 19622001. *Annals of Human Biology*. 36(5):584–594. DOI: 10.1080/03014460903136822.

Held, L. & Schwab, S. 2020. Improving the reproducibility of science. *Significance*. 17(1):10–11. DOI: 10.1111/j.1740-9713.2020.01351.x.

Henderson, C.Y. & Nikita, E. 2016. Accounting for multiple effects and the problem of small sample sizes in osteology: a case study focussing on entheseal changes. *Archaeological and Anthropological Sciences*. 8(4):805–817. DOI: 10.1007/s12520-015-0256-1.

Henn, B.M., Gignoux, C., Lin, A.A., Oefner, P.J., Shen, P., Scozzari, R., Cruciani, F., Tishkoff, S.A., et al. 2008. Y-chromosomal evidence of a pastoralist migration through Tanzania to southern Africa. *Proceedings of the National Academy of Sciences*. 105(31):10693–10698. DOI: 10.1073/pnas.0801184105.

Henneberg, M. & van den Berg, E.R. 1990. Test of socioeconomic causation of secular trend: Stature changes among favored and oppressed South Africans are parallel. *American Journal of Physical Anthropology*. 83(4):459–465. DOI: 10.1002/ajpa.1330830407.

Herbst, C.I., Tiemensma, M. & Wadee, S.A. 2015. A 10-year review of fatal community assault cases at a regional forensic pathology facility in Cape Town, South Africa. *South African Medical Journal*. 105(10):848–852. DOI: 10.7196/SAMJnew.8274.

Hernandez, C.J., Beaupré, G.S. & Carter, D.R. 2003. A theoretical analysis of the relative influences of peak BMD, age-related bone loss and menopause on the development of osteoporosis. *Osteoporosis International*. 14(10):843–847. DOI: 10.1007/s00198-003-1454-8.

Hochberg, Z. & Albertsson-Wikland, K. 2008. Evo-devo of infantile and childhood growth. *Pediatric Research*. 64(1):2–7. DOI: 10.1203/PDR.0b013e318177590f.

Hochberg, Z., Feil, R., Constancia, M., Fraga, M., Junien, C., Carel, J.C., Boileau, P., Le Bouc, Y., et al. 2011. DOI: 10.1210/er.2009-0039.

Hoskin, T. 2012. Parametric and nonparametric: demystifying the terms. *Ctsa.Mayo.Edu*. 1–5.

Humphrey, L.T. 1998. Growth patterns in the modern human skeleton. *American Journal of Physical Anthropology*. 105(1):57–72. DOI: 10.1002/(SICI)1096-8644(199801)105:1<57::AID-AJPA6>3.0.CO;2-A.

Huseynov, A., Zollikofer, C.P.E., Coudyzer, W., Gascho, D., Kellenberger, C., Hinzpeter, R. & De León, M.S.P. 2016. Developmental evidence for obstetric adaptation of the human female pelvis. *Proceedings of the National Academy of Sciences of the United States of America*. 113(19):5227–5232. DOI: 10.1073/pnas.1517085113.

Iguyon, I. & Elisseeff, A. 2003. An introduction to variable and feature selection. *Journal of Machine Learning Research*. 3:1157–1182. DOI: 10.1162/153244303322753616.

Inwood, K. & Masakure, O. 2013. Poverty and Physical Well-being among the Coloured Population in South Africa. *Economic History of Developing Regions*. 28(2):56–82. DOI: 10.1080/20780389.2013.866382.

Isaacs-Martin, W. & Petrus, T. 2012. The multiple meanings of coloured identity in South Africa. *Africa Insight*. 42(1):87–102.

Işcan, M.Y. & Steyn, M. 2013. Ancestry. In *The Human Skeleton in Forensic Medicine*. 3rd Edition Springfield, Illinois: Charles C Thomas Books. 195–226.

Jackson, D.A. & Chen, Y. 2004. Robust principal component analysis and outlier detection with ecological data. *Environmetrics*. 15(2):129–139. DOI: 10.1002/env.628.

James MacAluso, P. 2010. Sex determination from the acetabulum: Test of a possible non-population-specific discriminant function equation. *Journal of Forensic and Legal Medicine*. 17(6):348–351. DOI: 10.1016/j.jflm.2010.04.011.

Jayakumar, N., Athar, S. & Ashwood, N. 2020. Where do these cadavers come from? *Clinical Anatomy*. 33(6):872–875. DOI: 10.1002/ca.23570.

Jerković, I., Bašić, Ž., Anđelinović, Š. & Kružić, I. 2020. Adjusting posterior probabilities to meet predefined accuracy criteria: A proposal for a novel approach to osteometric sex estimation. *Forensic Science International*. 311(110273). DOI: 10.1016/j.forsciint.2020.110273.

Jonas. 2008. *Violin Plots for plotting multiple distributions (distributionPlot.m)*. Available: https://uk.mathworks.com/matlabcentral/fileexchange/23661-violin-plots-for-plotting-multiple-distributions-distributionplot-m [2021, November 06].

Jung, H.S. & Jung, H.S. 2009. Hand dominance and hand use behaviour reported in a survey of 2437 Koreans. *Ergonomics*. 52(11):1362–1371. DOI: 10.1080/00140130903067805.

Kanchan, T., Mohan Kumar, T.S., Pradeep Kumar, G. & Yoganarasimha, K. 2008. Skeletal asymmetry. *Journal of Forensic and Legal Medicine*. 15(3):177–179. DOI: 10.1016/j.jflm.2007.05.009.

Kelley, M.A. 1979. Sex Determination with Fragmented Skeletal Remains. *Journal of Forensic Sciences*. 24(1):10802J. DOI: 10.1520/jfs10802j.

Kemkes-Grottenthaler, A., Löbig, F. & Stock, F. 2002. Mandibular ramus flexure and gonial eversion as morphologic indicators of sex. *HOMO- Journal of Comparative Human Biology*. 53(2):97–111. DOI: 10.1078/0018-442X-00039.

Kidwell, P.A. 2015. Playing Checkers with Machines—from Ajeeb to Chinook. *Information & Culture*. 50(4):578–587. DOI: 10.7560/ic50405.

Kim, D.I., Kwak, D.S. & Han, S.H. 2013. Sex determination using discriminant analysis of the medial and lateral condyles of the femur in Koreans. *Forensic Science International*. 233(1–3):121–125. DOI: 10.1016/j.forsciint.2013.08.028.

Klales, A.R. 2016. Secular Change in Morphological Pelvic Traits used for Sex Estimation. *Journal of Forensic Sciences*. 61(2):295–301. DOI: 10.1111/1556-4029.13008.

Klaus, H.D. 2014. Frontiers in the bioarchaeology of stress and disease: Cross-disciplinary perspectives from pathophysiology, human biology, and epidemiology. *American Journal of Physical Anthropology*. 155(2):294–308. DOI: 10.1002/ajpa.22574.

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(1):77–90. DOI: 10.1002/ajpa.20934.

Konopka, T., Strona, M., Bolechała, F. & Kunz, J. 2007. Corpse dismemberment in the material collected by the Department of Forensic Medicine, Cracow, Poland. *Legal Medicine*. 9(1):1–13. DOI: 10.1016/j.legalmed.2006.08.008.

Kotěrová, A., Velemínská, J., Dupej, J., Brzobohatá, H., Pilný, A. & Brůžek, J. 2017. Disregarding population specificity: its influence on the sex assessment methods from the tibia. *International Journal of Legal Medicine*. 131(1):251–261. DOI: 10.1007/s00414-016-1413-5.

Kotsiantis, S.B. 2007. Supervised machine learning: A review of classification techniques. *Informatica*. 31(3):249–268. DOI: 10.31449/inf.v31i3.148.

Krishan, K., Chatterjee, P.M., Kanchan, T., Kaur, S., Baryah, N. & Singh, R.K. (in press). Forensic Anthropology Population Data, A review of sex estimation techniques during examination of skeletal remains in forensic anthropology casework. *Forensic Science International*. (261):165.e1-165.e8. DOI: 10.1016/j.forsciint.2016.02.007.

Krüger, G.C., L'Abbé, E.N. & Stull, K.E. 2017. Sex estimation from the long bones of modern South Africans. *International Journal of Legal Medicine*. 131(1):275–285. DOI: 10.1007/s00414-016-1488-z.

Kurki, H.K. & Decrausaz, S.L. 2016. Shape variation in the human pelvis and limb skeleton: Implications for obstetric adaptation. *American Journal of Physical Anthropology*. 159(4):630–638. DOI: 10.1002/ajpa.22922.

L'Abbe, E.N., Loots, M. & Meiring, J.H. 2005. The Pretoria Bone Collection: A modern South African skeletal sample. *Journal of Comparative Human Biology*. (56):197–205. DOI: 10.1016/j.jchb.2004.10.004.

Labuschagne, B.C.J. & Mathey, B. 2000. Cadaver profile at University of Stellenbosch Medical School, South Africa, 1956-1996. *Clinical Anatomy*. 13(2):88–93. DOI: 10.1002/(SICI)1098-2353(2000)13:2<88::AID-CA3>3.0.CO;2-Q.

Lambert, C.A. & Tishkoff, S.A. 2009. Genetic structure in African populations: Implications for human demographic history. In *Cold Spring Harbor Symposia on Quantitative Biology*. V. 74. NIH Public Access. 395–402. DOI: 10.1101/sqb.2009.74.053.

Latham, K.E. & Miller, J.J. 2019. DNA recovery and analysis from skeletal material in modern forensic contexts. *Forensic Sciences Research*. 4(1):51–59. DOI: 10.1080/20961790.2018.1515594.

Latimer, H.B. & Lowrance, E.W. 1965. Bilateral asymmetry in weight and in length of human bones. *The Anatomical Record*. 152(2):217–224. DOI: 10.1002/ar.1091520213.

Layden, E. 2019. *Simple Cohen's kappa*. Available: https://github.com/elayden/cohensKappa [2021, November 19].

Lazenby, R.A. 2002. Skeletal biology, functional asymmetry and the origins of "handedness." *Journal of Theoretical Biology*. 218(1):129–138. DOI: 10.1006/jtbi.2002.3052.

Liebenberg, L., L'Abbé, E.N. & Stull, K.E. 2015. Population differences in the postcrania of modern South Africans and the implications for ancestry estimation. *Forensic Science International*. 257:522–529. DOI: 10.1016/j.forsciint.2015.10.015.

Liebenberg, L., Krüger, G.C., L'Abbé, E.N. & Stull, K.E. 2019. Postcraniometric sex and ancestry estimation in South Africa: a validation study. *International Journal of Legal Medicine*. 133(1):289–296. DOI: 10.1007/s00414-018-1865-x.

Lilliefors, H.W. 1967. On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown. *Journal of the American Statistical Association*. 62(318):399–402. DOI: 10.1080/01621459.1967.10482916.

Lilliefors, H.W. 1969. On the Kolmogorov-Smirnov Test for the Exponential Distribution with Mean Unknown. *Journal of the American Statistical Association*. 64(325):387–389. DOI: 10.1080/01621459.1969.10500983.

Limony, Y., Friger, M. & Hochberg, Z. 2013. Pubertal gynecomastia coincides with peak height velocity. *JCRPE Journal of Clinical Research in Pediatric Endocrinology*. 5(3):142–144. DOI: 10.4274/Jcrpe.958.

Lin, L.I.-K. 1992. Assay Validation Using the Concordance Correlation Coefficient. *Biometrics*. 48(2):599. DOI: 10.2307/2532314.

Loth, S.R. & Henneberg, M. 1996. Mandibular ramus flexure: A new morphologic indicator of sexual dimorphism in the human skeleton. *American Journal of Physical Anthropology*. 99(3):473–485. DOI: 10.1002/(SICI)1096-8644(199603)99:3<473::AID-AJPA8>3.0.CO;2-X.

Loth, S.R. & Işcan, M.Y. 2000. Sex Determination. In *Encyclopedia of Forensic Sciences*. Sandiago: Academic Press.

Lovell, N.C. 1989. Test of Phenice's Technique for Determining Sex From the 0 s Pubis. *American Journal of Physical Anthropology*. 120:13–15.

Lundy, J.K. 1998. Forensic anthropology: What bones can tell us. *Laboratory Medicine*. 29(7):423–427. DOI: 10.1093/labmed/29.7.423.

Maass, P. & Friedling, L.J. (in press). Documented composition of cadaveric skeletal remains in the University of Cape Town Human Skeletal Collection, South Africa. *Forensic Science International*. 294:219.e1-219.e7. DOI: 10.1016/j.forsciint.2018.10.007.

Maass, P. & Friedling, L.J. 2019. Morphometric Analysis of the Neurocranium in an Adult South African Cadaveric Sample. *Journal of Forensic Sciences*. 64(2):367–374. DOI: 10.1111/1556-4029.13878.

MacAluso, P.J. 2011. Sex discrimination from the glenoid cavity in black South Africans: Morphometric analysis of digital photographs. *International Journal of Legal Medicine*. 125(6):773–778. DOI: 10.1007/s00414-010-0508-7.

Macho, G.A. 1990. Is sexual dimorphism in the femur a "population specific phenomenon." *Zeitschrift für Morphologie und Anthropologie*. 78(2):229–242.

Makgahlela, M., Sodi, T., Nkoana, S. & Mokwena, J. 2021. Bereavement rituals and their related psychosocial functions in a Northern Sotho community of South Africa. *Death Studies*. 45(2):91–100. DOI: 10.1080/07481187.2019.1616852.

Marsaglia, G., Tsang, W.W. & Wang, J. 2003. Evaluating Kolmogorov's Distribution. *Journal of Statistical Software*. 8(18).

Massey, F.J. 1951. The Kolmogorov-Smirnov Test for Goodness of Fit. *Journal of the American Statistical Association*. 46:68–78. DOI: 10.1080/01621459.1951.10500769.

MathWorks. 2006a. *Visualize summary statistics with box plot - MATLAB boxplot - MathWorks United Kingdom*. Available: https://uk.mathworks.com/help/stats/boxplot.html [2021, November 13].

MathWorks. 2006b. *Linear or rank correlation - MATLAB corr* . Available: https://uk.mathworks.com/help/stats/corr.html#d123e245452 [2021, November 15].

MathWorks. 2007. *Create Gaussian mixture model - MATLAB gmdistribution*. Available: https://uk.mathworks.com/help/stats/gmdistribution.html [2021, November 19].

MathWorks. 2008. *Partition data for cross-validation - MATLAB cvpartition*. Available: https://uk.mathworks.com/help/stats/cvpartition.html [2021, November 16].

MathWorks. 2012. *Cross-validate machine learning model - MATLAB crossval* . Available: https://uk.mathworks.com/help/stats/classificationsvm.crossval.html?searchHighlight=crossval&s_tid=srchtitle_crossval_2 [2021, November 24].

MathWorks. 2013. *Kernel Distribution*. Available: https://uk.mathworks.com/help/stats/kernel-distribution.html [2021, November 18].

MathWorks. 2014a. *Train multiclass naive Bayes model - MATLAB fitcnb*. Available: https://uk.mathworks.com/help/stats/fitcnb.html?searchHighlight=fitcnb&s_tid=srchtitle_fitcnb_1 [2021, November 18].

MathWorks. 2014b. *Fit k-nearest neighbor classifier - MATLAB fitcknn*. Available: https://uk.mathworks.com/help/stats/fitcknn.html [2021, November 18].

MathWorks. 2014c. *Fit binary decision tree for multiclass classification - MATLAB fitctree*. Available: https://uk.mathworks.com/help/stats/fitctree.html [2021, November 18].

MathWorks. 2014d. *Prediction Using Discriminant Analysis Models - MATLAB predict*. Available: https://uk.mathworks.com/help/stats/prediction-using-discriminant-analysis-models.html [2021, November 19].

MathWorks. 2014e. *Fit discriminant analysis classifier - MATLAB fitcdiscr*. Available: https://uk.mathworks.com/help/stats/fitcdiscr.html [2021, November 19].

MathWorks. 2016. *Fit ensemble of learners for classification - MATLAB fitcensemble*. Available: https://uk.mathworks.com/help/stats/fitcensemble.html [2021, November 19].

MathWorks. 2018. *Ensemble Algorithms*. Available: https://uk.mathworks.com/help/stats/ensemble-algorithms.html [2021, October 29].

MathWorks. 2019. *Rank features for classification using minimum redundancy maximum relevance (MRMR) algorithm - MATLAB fscmrmr*. Available: https://uk.mathworks.com/help/stats/fscmrmr.html [2021, November 21].

MathWorks. 2021a. *One-sample Kolmogorov-Smirnov test - MATLAB kstest*. Available: https://uk.mathworks.com/help/stats/kstest.html [2021, November 05].

MathWorks. 2021b. *Jarque-Bera test - MATLAB jbtest*. Available: https://uk.mathworks.com/help/stats/jbtest.html [2021, November 05].

MathWorks. 2021c. *Wilcoxon signed Rank test - MATLAB signrank*. Available: https://uk.mathworks.com/help/stats/signrank.html [2021, November 06].

MathWorks. 2021d. *Wilcoxon rank sum test - MATLAB ranksum*. Available: https://uk.mathworks.com/help/stats/ranksum.html [2021, November 06].

MathWorks. 2021e. *Naive Bayes classification*. Available: https://uk.mathworks.com/help/stats/naive-bayes-classification.html?s_tid=srchtitle [2021, October 29].

MathWorks. 2021f. *Cross-Validation - MATLAB & Simulink*. Available: https://uk.mathworks.com/discovery/cross-validation.html [2021, November 16].

Matthew, R. 2021. *f_CCC*. Available: https://github.com/robertpetermatthew/f_CCC [2021, August 10].

McHugh, M.L. 2012. Interrater reliability: The kappa statistic. *Biochemia Medica*. 22(3):276–282. DOI: 10.11613/bm.2012.031.

McNutt, M. 2014. Reproducibility. *Science*. 343(6168):229. DOI: 10.1126/SCIENCE.1250475/ASSET/42842A44-F51D-4427-A0F6-B26C842389C7/ASSETS/GRAPHIC/343_229_F2.JPEG.

Megyesi, M.S., Nawrocki, S.P. & Haskell, N.H. 2005. Using accumulated degree-days to estimate the postmortem interval from decomposed human remains. *Journal of forensic sciences*. 50(3):618–26.

Miller, L.H. 1956. Table of Percentage Points of Kolmogorov Statistics. *Journal of the American Statistical Association*. 51(273):111–121. DOI: 10.1080/01621459.1956.10501314.

Mircea, I.G. 2016. A fuzzy decision tree based method for skeletal sex determination. In *SACI 2016 - 11th IEEE International Symposium on Applied Computational Intelligence and Informatics, Proceedings*. Timişoara, Romania: IEEE. 447–452. DOI: 10.1109/SACI.2016.7507418.

Mishra, P., Pandey, C.M., Singh, U., Gupta, A., Sahu, C. & Keshri, A. 2019. Descriptive statistics and normality tests for statistical data. *Annals of Cardiac Anaesthesia*. 22(1):67–72. DOI: 10.4103/aca.ACA_157_18.

Mohd Razali, N. & Bee Wah, Y. 2011. Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *Journal of Statistical Modeling and Analytics*. 2(1):13–14.

Mokoena, P., Billings, B.K., Gibbon, V., Bidmos, M.A. & Mazengenya, P. 2019. Development of discriminant functions to estimate sex in upper limb bones for mixed ancestry South Africans. *Science and Justice*. 59(6):660–666. DOI: 10.1016/j.scijus.2019.06.007.

Montinaro, F., Busby, G.B.J., Gonzalez-Santos, M., Oosthuitzen, O., Oosthuitzen, E., Anagnostou, P., Destro-Bisol, G., Pascali, V.L., et al. 2017. Complex Ancient Genetic Structure and Cultural Transitions in Southern African Populations. *Genetics*. 205(1):303–316. DOI: 10.1534/genetics.116.189209.

Navega, D., Vicente, R., Vieira, D.N., Ross, A.H. & Cunha, E. 2015. Sex estimation from the tarsal bones in a Portuguese sample: a machine learning approach. *International Journal of Legal Medicine*. (129):651–659. DOI: 10.1007/s00414-014-1070-5.

Nikitovic, D. 2018. Sexual dimorphism (humans). In *The International Encyclopedia of Biological Anthropology*. Hoboken, NJ, USA: John Wiley & Sons, Inc. 1–4. DOI: 10.1002/9781118584538.ieba0443.

Oettlé, A.C., Pretorius, E. & Steyn, M. 2005. Geometric morphometric analysis of mandibular ramus flexure. *American Journal of Physical Anthropology*. 128(3):623–629. DOI: 10.1002/ajpa.20207.

Osisanwo, F.Y., Akinsola, J.E.T., Awodele, O., Hinmikaiye, J.O., Olakanmi, O. & Akinjobi, J. 2017. Supervised Machine Learning Algorithms: Classification and Comparison. *International Journal of Computer Trends and Technology*. 48(3):128–138. DOI: 10.14445/22312803/ijctt-v48p126.

Ousley, S. & Jantz, · R. 2013. Fordisc 3 Third generation of computer-aided forensic anthropology. DOI: 10.1007/s00194-013-0874-9.

Owings, W.P.A. & Myers, S.J. 2005. Epiphyseal union of the anterior iliac crest and medial clavicle in a modern multiracial sample of American males and females. *American Journal of Physical Anthropology*. 68(4):457–466. DOI: 10.1002/ajpa.1330680402.

Paluszek, M. & Thomas, S. 2017. *MATLAB Machine Learning*. S. Anglin, Ed. Berkeley, CA: Apress. DOI: 10.1007/978-1-4842-2250-8.

Patel, S., Shah, M., Vora, R., Goda, J., Rathod, S. & Shah, S. 2013. Morphometric analysis of scapula to determine sexual dimorphism. *International Journal of Medicine and Public Health*. 3(3):207. DOI: 10.4103/2230-8598.118946.

Patriquin, M.L., Loth, S.R. & Steyn, M. 2003. Sexually dimorphic pelvic morphology in South African whites and blacks. *HOMO- Journal of Comparative Human Biology*. 53(3):255–262. DOI: 10.1078/0018-442X-00049.

Patriquin, M.L., Steyn, M. & Loth, S.R. 2005. Metric analysis of sex differences in South African black and white pelves. *Forensic Science International*. 147(2–3):119–127. DOI: 10.1016/j.forsciint.2004.09.074.

Petersen, D.C., Libiger, O., Tindall, E.A., Hardie, R.A., Hannick, L.I., Glashoff, R.H., Mukerji, M., Fernandez, P., et al. 2013. Complex Patterns of Genomic Admixture within Southern Africa. *PLoS Genetics*. 9(3):e1003309. DOI: 10.1371/journal.pgen.1003309.

Phenice, T.W. 1969. A newly developed visual method of sexing the os pubis. *American Journal of Physical Anthropology*. 30(2):297–301. DOI: 10.1002/ajpa.1330300214.

Pillay, U. & Kramer, B. 1997. A simple method for the determination of sex from the pulp of freshly extracted human teeth utilizing the polymerase chain reaction. *The Journal of the Dental Association of South Africa = Die Tydskrif van die Tandheelkundige Vereniging van Suid-Afrika*. 52(11):673–677. Available: http://www.ncbi.nlm.nih.gov/pubmed/9589006 [2019, July 24].

Pillay, P., McQuoid Mason, D. & Satyapal, K.S. 2017. A study of the role and functions of inspectors of anatomy in South Africa. *South African Journal of Bioethics and Law*. 10(2):86. DOI: 10.7196/sajbl.2017.v10i2.00619.

Plochocki, J.H. 2004. Bilateral variation in limb articular surface dimensions. *American Journal of Human Biology*. 16(3):328–333. DOI: 10.1002/ajhb.20023.

Pretorius, E., Steyn, M. & Scholtz, Y. 2006. Investigation into the usability of geometric morphometric analysis in assessment of sexual dimorphism. *American Journal of Physical Anthropology*. 129(1):64–70. DOI: 10.1002/ajpa.20251.

Rawlik, K., Canela-Xandri, O. & Tenesa, A. 2016. Evidence for sex-specific genetic architectures across a spectrum of human complex traits. *Genome Biology*. 17(1):1–8. DOI: 10.1186/s13059-016-1025-x.

Recker, R.R., Davies, K.M., Hinders, S.M., Heaney, R.P., Stegman, M.R. & Kimmel, D.B. 1992. Bone Gain in Young Adult Women. *JAMA: The Journal of the American Medical Association*. 268(17):2403–2408. DOI: 10.1001/jama.1992.03490170075028.

Ren, Y., Zhang, L. & Suganthan, P.N. 2016. Ensemble Classification and Regression-Recent Developments, Applications and Future Directions [Review Article]. *IEEE Computational Intelligence Magazine*. 11(1):41–53. DOI: 10.1109/MCI.2015.2471235.

Ribot, I., Morris, A.G., Sealy, J. & Maggs, T. 2010. Population history and economic change in the last 2000 years in KwaZulu-Natal, RSA. *Southern African Humanities*. 22(1):89–112.

Robinson, M.S. & Bidmos, M.A. (in press). The skull and humerus in the determination of sex: Reliability of discriminant function equations. *Forensic Science International*. 186(1–3):86.e1-86.e5. DOI: 10.1016/j.forsciint.2009.01.003.

Roelofs, R., Miller, J., Hardt, M., Fridovich-keil, S., Schmidt, L. & Recht, B. 2019. A Meta-Analysis of Overfitting in Machine Learning. In *Conference on Neural Information Processing Systems*. Vancouver.

Rokade, S.A. & Gaikawad, A.P. 2012. Body donation in India: Social awareness, willingness, and associated factors. *Anatomical Sciences Education*. 5(2):83–89. DOI: 10.1002/ase.1263.

Ross, A.H. & Pilloud, M. 2021. The need to incorporate human variation and evolutionary theory in forensic anthropology: A call for reform. *American Journal of Physical Anthropology*. 176(4):672–683. DOI: 10.1002/ajpa.24384.

Scholtz, Y., Steyn, M. & Pretorius, E. 2010. A geometric morphometric study into the sexual dimorphism of the human scapula. *HOMO- Journal of Comparative Human Biology*. 61(4):253–270. DOI: 10.1016/j.jchb.2010.01.048.

Scott, M.A. 2019. Assessment of South African Metric Sex Estimation Methods using Long Bones and their Associated Material when Applied to a Mixed South African Sample. University of Cape Town.

Seedat, M., Niekerk, A. van, Jewkes, R., Suffla, S. & Ratele, K. 2009. Health in South Africa 5 Violence and injuries in South Africa: prioritising an agenda. *The Lancet*. 374(9694):1011–1022. DOI: 10.1016/S0140-6736(09)60948-X.

Seedat, M., van Niekerk, A., Jewkes, R., Suffla, S. & Ratele, K. 2009a. Violence and injuries in South Africa: prioritising an agenda for prevention. *The Lancet*. 374(9694):1011–1022. DOI: 10.1016/S0140-6736(09)60948-X.

Seedat, M., van Niekerk, A., Jewkes, R., Suffla, S. & Ratele, K. 2009b. DOI: 10.1016/S0140-6736(09)60948-X.

Seeman, E. 2003. The structural and biomechanical basis of the gain and loss of bone strength in women and men. *Endocrinology and Metabolism Clinics of North America*. 32(1):25–38. DOI: 10.1016/S0889-8529(02)00078-6.

Sen, P.C., Hajra, M. & Ghosh, M. 2020. Supervised Classification Algorithms in Machine Learning: A Survey and Review. In *Emerging Technology in Modelling and Graphics. Advances in Intelligent Systems and Computing, vol 937*. J. Mandal & D. Bhattacharya, Eds. Singapore: Springer Nature. 99–111. DOI: 10.1007/978-981-13-7403-6_11.

Shapiro, S.S. & Francia, R.S. 1972. An approximate analysis of variance test for normality. *Journal of the American Statistical Association*. 67(337):215–216. DOI: 10.1080/01621459.1972.10481232.

Shapiro, S.S. & Wilk, M.B. 1965. An Analysis of Variance Test for Normality (Complete Samples). *Biometrika*. 52(3/4):591. DOI: 10.2307/2333709.

Shields, A.E., Fortun, M., Hammonds, E.M., King, P.A., Lerman, C., Rapp, R. & Sullivan, P.F. 2005. The use of race variables in genetic studies of complex traits and the goal of reducing health disparities a transdisciplinary perspective. *American Psychologist*. 60(1):77–103. DOI: 10.1037/0003-066X.60.1.77.

da Silva, R. 2006. Creation of a database and a demographic survey of all cadavers used at the University of Cape Town Medical School 1911-2005. University of Cape Town. DOI: 10.4324/9780203334973_chapter_8.

Silventoinen, K., Sammalisto, S., Perola, M., Boomsma, D.I., Cornes, B.K., Davis, C., Dunkel, L., de Lange, M., et al. 2003. Heritability of Adult Body Height: A Comparative Study of Twin Cohorts in Eight Countries. *Twin Research*. 6(05):399–408. DOI: 10.1375/twin.6.5.399.

Simmons, T., Jantz, R.L. & Bass, W.M. 2015. Stature Estimation from Fragmentary Femora: A Revision of the Steele Method. *Journal of Forensic Sciences*. 35(3):628–636. DOI: 10.1520/jfs12868j.

Skoglund, P., Thompson, J.C., Prendergast, M.E., Mittnik, A., Sirak, K., Hajdinjak, M., Salie, T., Rohland, N., et al. 2017. Reconstructing Prehistoric African Population Structure. *Cell*. 171(1):59-71.e21. DOI: 10.1016/j.cell.2017.08.049.

Šlaus, M., Bedić, Ž., Strinović, D. & Petrovečki, V. (in press). Sex determination by discriminant function analysis of the tibia for contemporary Croats. *Forensic Science International*. 226(1–3):302.e1-302.e4. DOI: 10.1016/j.forsciint.2013.01.025.

Smith, R.W. & Walker, R.R. 1964. Femoral Expansion in Aging Women: Implications for Osteoporosis and Fractures. *Science*. 145(3628):156–157. DOI: 10.1126/science.145.3628.156.

Sommer, C. & Gerlich, D.W. 2013. Machine learning in cell biology-teaching computers to recognize phenotypes. *Journal of Cell Science*. 126(24):5529–5539. DOI: 10.1242/jcs.123604.

Soni, G., Dhall, U. & Chhabra, S. 2010. Determination of sex from femur: Discriminant analysis. *Journal of the Anatomical Society of India*. 59(2):216–221. DOI: 10.1016/S0003-2778(10)80029-2.

Spennemann, D.H.R. & Franke, B. 1995. Decomposition of Buried Human Bodies and Associated Death Scene Materials on Coral Atolls in the Tropical Pacific. *Journal of Forensic Sciences*. 40(3):13787J. DOI: 10.1520/jfs13787j.

Spies, M.J., Gibbon, V.E. & Finaughty, D.A. 2018. Forensic taphonomy: Vertebrate scavenging in the temperate southwestern Cape, South Africa. *Forensic Science International*. 290(2018):62–69. DOI: 10.1016/j.forsciint.2018.06.022.

Spies, M.J., Finaughty, D.A. & Gibbon, V.E. 2018. Forensic taphonomy: Scavenger-induced scattering patterns in the temperate southwestern Cape, South Africa — A first look. *Forensic Science International*. 290(2018):29–35. DOI: 10.1016/j.forsciint.2018.06.015.

Spradley, M.K. & Jantz, R.L. 2011. Sex estimation in forensic anthropology: skull versus postcranial elements. *Journal of forensic sciences*. 56(2):289–96. DOI: 10.1111/j.1556-4029.2010.01635.x.

Statistics South Africa. 2018. *Mortality and causes of death in South Africa: Findings from death notification*. Pretoria.

Statistics South Africa. 2019. *Mid-year population estimates*. Pretoria. Available: www.statssa.gov.za,info@statssa.gov.za,Tel+27123108911 [2021, January 27].

Steinskog, D.J., Tjøtheim, D.B. & Kvamstø, N.G. 2007. A cautionary note on the use of the Kolmogorov-Smirnov test for normality. *Monthly Weather Review*. 135(3):1151–1157. DOI: 10.1175/MWR3326.1.

Steyn, M. & Işcan, M.Y. 1997. Sex determination from the femur and tibia in South African whites. *Forensic Science International*. 90(1–2):111–119. DOI: 10.1016/S0379-0738(97)00156-4.

Steyn, M. & Işcan, M.Y. 1999. Osteometric variation in the humerus: Sexual dimorphism in South Africans. *Forensic Science International*. 106(2):77–85. DOI: 10.1016/S0379-0738(99)00141-3.

Steyn, M. & Patriquin, M.L. (in press). Osteometric sex determination from the pelvis-Does population specificity matter? *Forensic Science International*. 191(1–3):113.e1-113.e5. DOI: 10.1016/j.forsciint.2009.07.009.

Steyn, M., Meiring, J.H. & Nienaber, W.C. 1997. Forensic anthropology in South Africa: a profile of cases from 1993 to 1995 at the DepartJ.11ent of Anatomy, University of Pretoria. *South African Journal of Ethnology*. 20(1):23–26.

Steyn, M., Pretorius, E. & Hutten, L. 2004. Geometric morphometric analysis of the greater sciatic notch in South Africans. *HOMO- Journal of Comparative Human Biology*. 54(3):197–206. DOI: 10.1078/0018-442X-00076.

Stojanowski, C.M., Seidemann, R.M. & Doran, G.H. 2002. Differential skeletal preservation at Windover Pond: Causes and consequences. *American Journal of Physical Anthropology*. 119(1):15–26. DOI: 10.1002/ajpa.10101.

Stulp, G. & Barrett, L. 2016. Evolutionary perspectives on human height variation. *Biological Reviews*. 91(1):206–234. DOI: 10.1111/brv.12165.

Sutherland, L.D. & Suchey, J.M. 1991. Use of the ventral arc in pubic sex determination. *Journal of forensic sciences*. 36(2):501–11. Available: http://www.ncbi.nlm.nih.gov/pubmed/2066725.

Szulc, P. 2006. Bone density, geometry, and fracture in elderly men. *Current Osteoporosis Reports*. 4(2):57–63. DOI: 10.1007/s11914-006-0003-8.

Tamir, M. 2020. *What is Machine Learning?* Available: https://ischoolonline.berkeley.edu/blog/what-is-machine-learning/ [2021, October 21].

Tawha, T., Dinkele, E., Mole, C. & Gibbon, V.E. 2020. Assessing zygomatic shape and size for estimating sex and ancestry in a South African sample. *Science and Justice*. 60(3):284–292. DOI: 10.1016/j.scijus.2020.01.003.

Thadewald, T. & Büning, H. 2007. Jarque-Bera test and its competitors for testing normality - A power comparison. *Journal of Applied Statistics*. 34(1):87–105. DOI: 10.1080/02664760600994539.

Toneva, D.H., Nikolova, S.Y., Agre, G.P., Zlatareva, D.K., Hadjidekov, V.G. & Lazarov, N.E. 2020. Data mining for sex estimation based on cranial measurements. *Forensic Science International*. 315(110441). DOI: 10.1016/j.forsciint.2020.110441.

Urbani, C., Lastrucci, R.D. & Kramer, B. 1999. The effect of temperature on sex determination using DNA-PCR analysis of dental pulp. *Journal of Forensic Odonto-Stomatology*. 17(2):35–39. Available: http://www.ncbi.nlm.nih.gov/pubmed/10709561 [2019, July 24].

Vabalas, A., Gowen, E., Poliakoff, E. & Casson, A.J. 2019. Machine learning algorithm validation with a limited sample size. *PLoS ONE*. 14(11):e0224365. DOI: 10.1371/journal.pone.0224365.

Vance, V.L. 2007. Age Related Changes in the Post Cranial Huma Skeleton and its Implication for the Determination of Sex. PhD Anatomy. University of Pretoria.

Vance, V.L. & Steyn, M. 2013. Geometric morphometric assessment of sexually dimorphic characteristics of the distal humerus. *HOMO- Journal of Comparative Human Biology*. 64(5):329–340. DOI: 10.1016/j.jchb.2013.04.003.

Vance, V.L., Steyn, M. & L'Abbé, E.N. 2011. Nonmetric sex determination from the distal and posterior humerus in black and white South Africans. *Journal of Forensic Sciences*. 56(3):710–714. DOI: 10.1111/j.1556-4029.2011.01724.x.

Victora, C.G., Adair, L., Fall, C., Hallal, P.C., Martorell, R., Richter, L. & Sachdev, H.S. 2008. Maternal and child undernutrition: consequences for adult health and human capital. *The Lancet*. 371(9609):340–357. DOI: 10.1016/S0140-6736(07)61692-4.

De Villiers, H. 1968. Sexual dimorphism of the skull of the South African Banu-speaking Negro. *South African Journal of Science*. 64(2):118–124.

Walker, P.L. 2005. Greater sciatic notch morphology: Sex, age, and population differences. *American Journal of Physical Anthropology*. 127(4):385–391. DOI: 10.1002/ajpa.10422.

Walrath, D.E., Turner, P. & Bruzek, J. 2004. Reliability test of the visual assessment of cranial traits for sex determination. *American Journal of Physical Anthropology*. 125(2):132–137. DOI: 10.1002/ajpa.10373.

Weedon, M.N. & Frayling, T.M. 2008. Reaching new heights: insights into the genetics of human stature. *Trends in Genetics*. 24(12):595–603. DOI: 10.1016/j.tig.2008.09.006.

Wells, J.C.K. 2007. Sexual dimorphism of body composition. *Best Practice and Research Clinical Endocrinology and Metabolism*. 21(3):415–430. DOI: 10.1016/j.beem.2007.04.007.

Wells, J.C.K. 2012. Ecogeographical associations between climate and human body composition: Analyses based on anthropometry and skinfolds. *American Journal of Physical Anthropology*. 147(2):169–186. DOI: 10.1002/ajpa.21591.

Willey, P., Galloway, A. & Snyder, L. 1997. Bone mineral density and survival of elements and element portions in the bones of the Crow Creek massacre victims. *American Journal of Physical Anthropology*. 104(4):513–528. DOI: 10.1002/(SICI)1096-8644(199712)104:4<513::AID-AJPA6>3.0.CO;2-S.

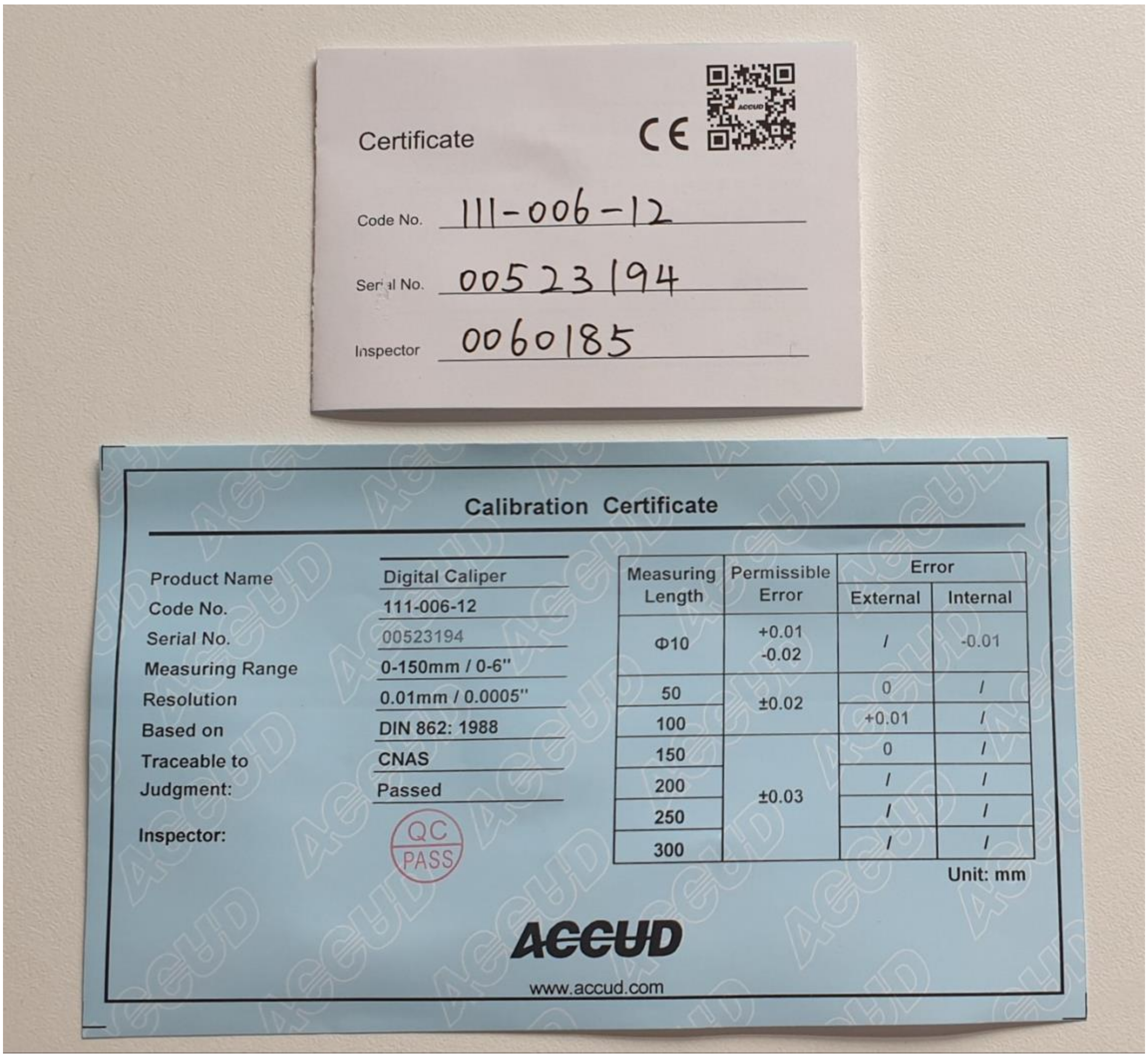
De Wit, E., Delport, W., Rugamika, C.E., Meintjes, A., Möller, M., Helden, P.D. Van, Seoighe, C. & Hoal, E.G. 2010. Genome-wide analysis of the structure of the South African Coloured Population in the Western Cape. *Human Genetics*. 128:145–153. DOI: 10.1007/s00439-010-0836-1.

Woodward, V.E., Penny, C.B. & Ruff, P. 2006. Intercondylar Fossa of the Femur: A Novel Region for DNA Extraction. *South African Archaeological Society*. 61(183):96–97.

Zulu, E. 2013. Reverence For Ancestors in Africa: Interpretation Of The 5th Commandment From An African Perspective. *Scriptura*. 81:476. DOI: 10.7833/81-0-749.

APPENDICES

APPENDIX A: CALIBRATION CERTIFICATE



Certificate

CE

Code No. 111-006-12

Serial No. 00523194

Inspector 0060185

Calibration Certificate

Product Name	Digital Caliper
Code No.	111-006-12
Serial No.	00523194
Measuring Range	0-150mm / 0-6"
Resolution	0.01mm / 0.0005"
Based on	DIN 862: 1988
Traceable to	CNAS
Judgment:	Passed
Inspector:	QC PASS

Measuring Length	Permissible Error	Error	
		External	Internal
Φ10	+0.01 -0.02	/	-0.01
50	±0.02	0	/
100		+0.01	/
150	±0.03	0	/
200		/	/
250		/	/
300		/	/

Unit: mm

ACCUD

www.accud.com

Appendix A1: Calibration certificate for digital sliding callipers **used** *for data collection.*

APPENDIX B: FACTORS WHICH INFLUENCE REPLICABILITY AND REPRODUCIBILITY OF ML MODELLING

Several factors have a marked impact on the final model trained by an ML algorithm and thus have the capacity to influence replicability and reproducibility of results.

B1 DIFFERENCES CAUSED BY LEARNING ALGORITHM

One can use the same dataset to train the same algorithm multiple times and get a different resultant model every time. This is the nature of some machine learning algorithms. It is not an error or a ‘bug’ but rather a feature of some programs. This is undeniably a concern when it comes to reproducibility.

Some machine learning algorithms are *deterministic.* This means that, given the same dataset multiple times, an algorithm will learn the same model on every run. Examples of deterministic algorithms are discriminant function analysis and most decision trees. The benefit of deterministic algorithms is that they are easy to reproduce but, they are often far more simplistic and unable to discover complex underlying patterns in the data.

Some ml algorithms are not deterministic but rather incorporate elements of entropy or randomness. These algorithms are *stochastic.* This doesn’t mean that the algorithms are random but rather that the specific small decisions made during learning can vary randomly. The impact of this is that each time a stochastic learner is run on the same data, it learns a slightly different model and may make slightly different predictions which, in turn, may be detrimental to reproducibility.

The benefit of stochastic algorithms is that that they allow an algorithm to break symmetry and escape local optima in favour of global optima to find the best possible mapping of inputs to outputs. Bagging (random forest) is an example of a stochastic machine learning algorithm. Randomness is used in the sampling process to ensure that many different decision trees are grown to make up the random forest and stimulate independent predictions from the same dataset.

Randomness in algorithm training can, in part, be controlled by setting the seed for the random number generator (rng), which ensures that the same randomness is employed every time the algorithm is run. But what is the best seed? What's to say the chosen seed is even a good one?

The answer to the first question, we don't know what the best seed is. But for the sake of reproducibility the seed for the rng was always set before algorithm training and can be found in the analysis code (https://github.com/mscott1037/masters.). This means that running the code multiple times should, in theory, produce the same result.

To answer whether the seed is a good one? There are more than 4 billion (4^{33}-1, to be specific) possible values for the rng seed in MATLAB. It is simply not plausible to test all of these and unlikely that the vast majority of seed values will even have a significant impact on the final model. Common practice is to rerun one's code using different seed values a couple of times, apply new data to test all the models and report the average accuracy. The best way to incorporate this and embrace the randomness is by using ensemble methods but these are difficult to interpret and impossible to use without software. In principle, a very similar process is followed during k-fold cross validation where subsets of the training data are used to train and test models and the mean accuracy is reported. In addition, a *post hoc* sample size analysis was performed which enables one to determine what the expected level of variability is between trained models.

B2 DIFFERENCES CAUSED BY PLATFORM

Another important contributor to consider in assessing replicability and reproducibility when applying machine learning to a classification problem is hardware and software.

System architecture (whether your system runs tasks on a GPU or CPU), operating system (does your machine run Windows or MacOS), underlying maths libraries and software versions (i.e., all analysis for this study was performed in MATLAB version 2020b rather than updating to the most current version - MATLAB 2021b) play a role in the results of computational analyses.

Machine learning algorithms are complex numerical computations containing floating point values. Floating point values take numbers with a long string of decimals and select a set number of significant digits, scaled using an exponent (like scientific notation). These are called floating points because the number's decimal place can be positioned anywhere relative to the significant digits of the number. Differences in the computing system can result in different rounding of numbers which, when compounded over many computational steps, may lead to very different results.

In practice, the differences are likely very small but nonetheless the same 'platform' was used for all computational tasks, and it is acknowledged that this may contribute to differences in any future work which may attempt to replicate the model training in this study.

B3 DIFFERENCES CAUSED BY TRAINING DATA

Small differences in the training data can have large implications on the final trained model if the dataset is relatively small (Brain & Webb, 1999). All algorithms are sensitive to differences in the data – this is known as variance. Algorithms with high variance require larger sample sizes than those with lower variance as outliers and extreme values have less impact on model parameters when the dataset is large.

Decision trees are an example of a high variance classifier and as such larger sample sizes are optimal to reduce variance. If the sample size is small and one wishes to train a decision tree algorithm, ensemble methods like bagging and boosting can be employed to counteract variance (Dietterich & Kong, 1995).

APPENDIX C: SEX CLASSIFICATION MODELS

Note: Here we refer to a skewing of prediction accuracy towards one class (male or female) as a 'sex bias'. A difference of <2.0% (~10 individuals) or less will be considered insignificant, ~2.0-4.0% minimal, ~5.0-7.0% moderate and >8.0% (~40 individuals) large.

C1 UNIVARIATE MODELS

Univariate Naïve Bayes, KNN, decision tree, DFA and ensemble models were trained for each of the 12 univariate predictors in this study. Results pertaining to best fit models and most useful predictors are analysed in the sections that follow.

C1.1 ACETABULUM DIAMETER (AD) - TABLE C1-1

Accuracies for the different AD models are uniform, with less than a 1.0% *k*-fold CV accuracy difference between the highest and lowest result (Table C1-1) The ADknn, ADtree and ADensemble models have moderate to large sex bias in their *k*-fold CV accuracies. Taking sex bias and holdout kappa scores into consideration, the 2 best models are ADnb and ADdiscr. The more user-friendly discriminant model was thus selected as best fit.

C1.2 TRANSVERSE ACETABULUM DIAMETER (TAD) – TABLE C1-2

TAD uses the same sample as AD but k-fold CV accuracies for TAD are on average between 1.0% and 2.0% lower than for AD. The k-fold CV accuracies for TAD range from 82.2% to 83.8% (slightly lower than but very similar to training accuracies). TADnb model is the worst fit of all TAD models. All models show moderate to large sex biases under k-fold CV except the discriminant model, which is thus selected as the best fit model.

C1.3 FEMORAL HEAD DIAMETER (FHD) - TABLE C1-3

Accuracy results between all models are once again very similar. FHDtree model had the lowest *k*-fold CV accuracy result (85.2%). The FHDnb and FHDdiscr models have joint highest *k*-fold CV accuracy and kappa statistic (0.74) results. The more user-friendly discriminant model was thus selected as best fit.

Table C1-1 Goodness of fit for all univariate AD models.

	Training Accuracy				***K*-fold Accuracy**				**Holdout Accuracy**			
Model	All	Male	Female	Kappa	All	Male	Female	Kappa	All	Male	Female	Kappa
ADnb	84.58	83.13	85.99	0.69	84.78	83.13	86.38	0.70	85.33	90.16	80.65	0.71
ADknn	85.18	83.13	87.16	0.70	84.78	81.93	87.55	0.70	83.72	86.89	80.65	0.67
ADtree	85.57	79.92	91.05	0.71	84.98	79.92	89.88	0.70	84.55	85.25	83.87	0.69
ADdiscr	84.39	83.94	84.82	0.69	84.39	83.94	84.82	0.69	86.14	91.80	80.65	0.72
ADensemble	85.57	79.92	91.05	0.71	84.78	79.92	89.49	0.70	84.55	85.25	83.87	0.69

Table C1-2 Goodness of fit for all univariate TAD models.

	Training Accuracy				***K*-fold Accuracy**				**Holdout Accuracy**			
Model	All	Male	Female	Kappa	All	Male	Female	Kappa	All	Male	Female	Kappa
TADnb	82.21	73.49	90.66	0.64	82.21	73.09	91.05	0.64	83.77	80.33	87.10	0.67
TADknn	83.20	79.12	87.16	0.66	82.81	79.12	86.38	0.66	85.35	86.89	83.87	0.71
TADtree	84.19	77.91	90.27	0.68	83.79	77.51	89.88	0.68	83.75	81.97	85.48	0.67
TADdiscr	82.61	83.53	81.71	0.65	82.81	83.53	82.10	0.66	85.33	90.16	80.65	0.71
TADensemble	84.19	77.91	90.27	0.68	83.79	77.51	89.88	0.68	83.75	81.97	85.48	0.67

Table C1-3 Goodness of fit for all univariate FHD models.

	Training Accuracy				***K*-fold Accuracy**				**Holdout Accuracy**			
Model	All	Male	Female	Kappa	All	Male	Female	Kappa	All	Male	Female	Kappa
FHDnb	87.32	86.25	88.35	0.75	86.91	86.25	87.55	0.74	82.31	86.44	78.33	0.65
FHDknn	86.09	85.83	86.35	0.72	86.50	85.83	87.15	0.73	81.48	84.75	78.33	0.63
FHDtree	87.32	85.42	89.16	0.75	85.28	85.83	85.54	0.71	82.33	84.75	80.00	0.65
FHDdiscr	87.12	86.25	87.95	0.74	86.91	86.25	87.55	0.74	82.31	86.44	78.33	0.65
FHDensemble	87.32	85.83	88.76	0.75	86.50	85.83	87.15	0.73	82.33	84.75	80.00	0.65

C1.4 TRANSVERSE FEMORAL HEAD DIAMETER (THD) – TABLE C1-4

Results for THD are very similar to FHD. Once again, there is very little difference in *k*-fold CV accuracy between the 5 models which have accuracies ranging from 86.1-86.9%. The best fit models are the THDnb and THDdiscr models. The discriminant model is superior as it presents a lower *k*-fold CV sex bias (1.0%) than the NB model (2.9%).

C1.5 MEDIAL CONDYLAR LENGTH (MCL) - TABLE C1-5

MCL models are the second least accurate of all univariate models with *k*-fold CV accuracies between 80.1-81.2%. All models except MCLdiscr exhibit a large sex bias in *k*-fold CV accuracy scores. The discriminant model has the highest holdout accuracy and the lowest *k*-fold CV sex bias and is thus selected as the best fit model.

C1.6 BICONDYLAR BREADTH (BB) - TABLE C1-6

All BB models share similar *k*-fold CV accuracy scores (83.4-84.2%). The tree and ensemble models have a *k*-fold CV sex bias of more than 10.0% and are thus not optimal. The BBdiscr model shows no *k*-fold CV sex bias (<1% difference) and an 83.4% *k*-fold CV accuracy and is thus chosen as the most reliable and accurate BB model.

Table C1-4 Goodness of fit for all univariate THD models.

	Training Accuracy				***K*-fold Accuracy**				**Holdout Accuracy**			
Model	All	Male	Female	Kappa	All	Male	Female	Kappa	All	Male	Female	Kappa
THDnb	87.12	85.83	88.35	0.74	86.91	85.42	88.35	0.74	83.98	89.83	78.33	0.68
THDknn	87.12	85.42	88.76	0.74	86.71	84.58	88.76	0.73	83.99	88.14	80.00	0.68
THDtree	87.32	83.75	90.76	0.75	86.09	85.00	86.75	0.72	83.16	86.44	80.00	0.66
THDdiscr	86.50	85.83	87.15	0.73	86.71	86.25	87.15	0.73	84.81	91.53	78.33	0.70
THDensemble	87.32	85.83	88.76	0.75	86.09	84.17	87.95	0.72	83.98	89.83	78.33	0.68

Table C1-5 Goodness of fit for all univariate MCL models.

	Training Accuracy				***K*-fold Accuracy**				**Holdout Accuracy**			
Model	All	Male	Female	Kappa	All	Male	Female	Kappa	All	Male	Female	Kappa
MCLnb	81.15	77.41	84.74	0.62	81.15	77.41	84.74	0.62	77.55	75.00	80.00	0.55
MCLknn	80.74	76.15	85.14	0.61	80.12	74.90	85.14	0.60	78.44	73.33	83.33	0.57
MCLtree	81.56	78.66	84.34	0.63	80.53	76.15	84.34	0.61	77.52	76.67	78.33	0.55
MCLdiscr	80.12	79.08	81.12	0.60	80.33	79.08	81.53	0.61	79.97	81.67	78.33	0.60
MCLensemble	81.56	78.66	84.34	0.63	80.12	75.73	84.34	0.60	77.52	76.67	78.33	0.55

Ensemble KNN?

Table C1-6 Goodness of fit for all univariate BB models.

	Training Accuracy				***K*-fold Accuracy**				**Holdout Accuracy**			
Model	All	Male	Female	Kappa	All	Male	Female	Kappa	All	Male	Female	Kappa
BBnb	84.02	82.43	85.54	0.68	84.02	82.43	85.54	0.68	83.30	85.00	81.67	0.67
BBknn	84.43	82.01	86.75	0.69	83.61	80.33	86.75	0.67	79.22	76.67	81.67	0.58
BBtree	84.63	73.64	95.18	0.69	84.22	73.64	94.78	0.69	77.76	65.00	90.00	0.55
BBdiscr	84.02	84.10	83.94	0.68	83.40	83.68	83.13	0.67	83.30	85.00	81.67	0.67
BBensemble	84.84	77.82	91.57	0.70	84.22	78.24	89.96	0.68	78.54	68.33	88.33	0.57

C1.7 LATERAL CONDYLAR LENGTH (LCL) - TABLE C1-7

LCL models are overall the least accurate models (k-fold CV kappa 0.6%). The LCLdiscr model performs best with the highest kappa statistics as well as training, k-fold CV and ho accuracies amongst all LCL models. MCLdiscr (80.3%) has a minorly lower k-fold CV accuracy than LCLdiscr (80.5%) both of which are selected as the best fit model for their respective predictors.

C1.8 TIBIA PROXIMAL BREADTH (TPB) - TABLE C1-8

TPB is the only tibial predictor and thus represents the bone model as well. Tibial models have relatively high *k*-fold CV kappa values and differ in *k*-fold CV accuracies by only ~1.0%. The highest *k*-fold CV kappa statistics belong the TibiaNb (0.76) and TibiaEnsemble (0.76) models. Both models however exhibit sex biases. The discriminant model is a little more than 1% less accurate under *k*-fold CV but the sex bias is much lower, so it surpasses the other models.

C1.9 GLENOID LENGTH (GL) - TABLE C1-9

GL models have the highest k-fold CV accuracies (88.3-89.3%) and kappa values (0.77-0.79) of all univariate models. Model accuracies are fairly uniform however, the GLtree and GLensemble models have the highest *k*-fold CV kappa statistic. Although accuracies for both of these models are incredibly similar, the tree model is superior as the small increase in accuracy between the 2 models is not large enough to justify the increased complexity of an ensemble model. GLdiscr is chosen as the best fit model as the *k*-fold CV accuracy is not much lower than GLtree but there is much less sex bias.

Table C1-7 Goodness of fit for all univariate LCL models.

	Training Accuracy				*K*-fold Accuracy				Holdout Accuracy			
Model	All	Male	Female	Kappa	All	Male	Female	Kappa	All	Male	Female	Kappa
LCLnb	80.53	74.90	85.94	0.61	80.12	74.48	85.54	0.60	75.14	68.33	81.67	0.50
LCLknn	80.12	73.64	86.35	0.60	79.92	74.90	84.74	0.60	75.14	68.33	81.67	0.50
LCLtree	80.94	72.80	88.76	0.62	78.69	71.13	88.76	0.60	75.17	66.67	83.33	0.50
LCLdiscr	80.94	79.08	82.73	0.62	80.53	78.66	82.33	0.61	75.88	73.33	78.33	0.52
LCLensemble	80.74	76.15	85.14	0.61	78.48	74.06	82.73	0.57	75.95	70.00	81.67	0.52

Table C1-8 Goodness of fit for all univariate tibia/TPB models.

	Training Accuracy				*K*-fold Accuracy				Holdout Accuracy			
Model	All	Male	Female	Kappa	All	Male	Female	Kappa	All	Male	Female	Kappa
TibiaNb	88.50	92.17	85.11	0.77	88.05	91.71	84.68	0.76	87.31	96.36	78.95	0.75
TibiaKnn	87.61	84.33	90.64	0.75	87.17	84.79	89.36	0.74	85.64	89.09	82.46	0.71
TibiaTree	88.72	91.71	85.96	0.77	86.28	91.71	83.40	0.75	87.31	96.36	78.95	0.75
TibiaDiscr	86.73	85.71	87.66	0.73	87.17	86.64	87.66	0.74	87.39	92.73	82.46	0.75
TibiaEnsemble	90.27	90.78	89.79	0.81	88.27	85.71	90.64	0.76	82.94	87.27	78.95	0.66

Table C1-9 Goodness of fit for all univariate GL models.

	Training Accuracy				*K*-fold Accuracy				Holdout Accuracy			
Model	All	Male	Female	Kappa	All	Male	Female	Kappa	All	Male	Female	Kappa
GLnb	89.88	91.39	88.40	0.80	88.87	90.16	87.60	0.78	84.16	85.00	83.33	0.68
GLknn	88.26	85.66	90.80	0.77	88.26	86.48	90.00	0.77	83.33	83.33	83.33	0.67
GLtree	89.88	92.21	87.60	0.80	89.07	91.39	87.20	0.79	84.14	86.67	81.67	0.68
GLdiscr	88.87	88.11	89.60	0.78	88.66	88.11	89.20	0.77	83.33	83.33	83.33	0.67
GLensemble	89.88	92.21	87.60	0.80	89.27	90.98	87.60	0.79	84.14	86.67	81.67	0.68

C1.10 GLENOID BREATH (GB) - TABLE C1-10

All GB models have very similar *k*-fold CV accuracy results (84.0-85.2%) and kappa statistics (0.68-0.70). The KNN model has the highest k-fold CV accuracy, joint highest *k*-fold CV kappa static and the least *k*-fold CV sex bias however, the KNN *K*-value = 87 (very large). The second highest *k*-fold CV accuracy and kappa statistic belongs to the GBnb model; however, the sex bias is large (8.9%). Thus, despite having a lower overall accuracy and kappa statistic, the GBdiscr model, with its minimal sex bias, outdoes all other GB models.

C1.11 EPICONDYLAR BREADTH (EB) - TABLE C1-11

Across all EB models, *k*-fold CV accuracies are very similar with barely more than 1.0% difference between the lowest and highest accuracies (86.9-87.9%). Despite being slightly inferior to EBnb in *k*-fold CV accuracy, the discriminant model shows significant *k*-fold CV sex bias and is thus selected as the bit fit model for EB.

C1.12 HUMERUS HEAD DIAMETER (HHD) - TABLE C1-12

The HHDknn model has a training accuracy of 99.2% and a *k*-fold CV accuracy of only 82.6%. This model is likely overfit. The HHDnb and HHDdiscr models have the best training, *k*-fold CV and ho kappa statistics. The ease of use for discriminant models over NB makes them the best choice of model for sex estimation when all other factors are equal.

C2 BONE MODELS

Multivariate bone-specific naïve Bayes, KNN, decision tree, discriminant and ensemble models were trained and optimised for the Pelvis (AD + TAD), Femur (FHD + THD + MCL + BB + LCL), Scapula (GL + GB) and, Humerus (EB + HHD). Results pertaining to best fit models and most useful models are set out in the sections to follow.

Table C1-10 Goodness of fit for all univariate GB models.

	Training Accuracy				***K*-fold Accuracy**				**Holdout Accuracy**			
Model	All	Male	Female	Kappa	All	Male	Female	Kappa	All	Male	Female	Kappa
GBnb	85.22	80.74	89.60	0.70	85.22	80.74	89.60	0.70	85.02	83.33	86.67	0.70
GBknn	85.43	83.20	87.60	0.71	85.22	84.84	85.60	0.70	85.84	85.00	86.67	0.72
GBtree	85.83	79.92	91.60	0.72	84.82	79.51	89.60	0.69	84.20	81.67	86.67	0.68
GBdiscr	84.62	86.48	82.80	0.69	84.62	86.48	82.80	0.69	88.29	91.67	85.00	0.77
GBensemble	85.43	81.56	89.20	0.71	84.01	79.92	88.00	0.68	85.84	85.00	86.67	0.72

Table C1-11 Goodness of fit for all univariate EB models.

	Training Accuracy				***K*-fold Accuracy**				**Holdout Accuracy**			
Model	All	Male	Female	Kappa	All	Male	Female	Kappa	All	Male	Female	Kappa
EBnb	87.93	89.45	86.51	0.76	87.93	89.45	86.51	0.76	83.41	90.16	77.05	0.67
EBknn	87.73	89.45	86.11	0.75	87.73	89.45	86.11	0.75	83.41	90.16	77.05	0.67
EBtree	88.14	90.30	86.11	0.76	86.91	90.30	84.92	0.75	83.41	90.16	77.05	0.67
EBdiscr	87.32	87.34	87.30	0.75	87.32	87.34	87.30	0.75	85.15	88.52	81.97	0.70
EBensemble	88.14	90.30	86.11	0.76	87.12	90.30	84.13	0.74	83.41	90.16	77.05	0.67

Table C1-12 Goodness of fit for all univariate HHD models.

	Training Accuracy				***K*-fold Accuracy**				**Holdout Accuracy**			
Model	All	Male	Female	Kappa	All	Male	Female	Kappa	All	Male	Female	Kappa
HHDnb	83.64	83.12	84.13	0.67	83.44	82.70	84.13	0.67	82.66	86.89	78.69	0.66
HHDknn	99.18	98.31	100.00	0.98	82.62	79.75	85.32	0.65	84.35	86.89	81.97	0.69
HHDtree	84.25	81.86	86.51	0.68	81.80	79.75	84.13	0.64	85.20	86.89	83.61	0.70
HHDdiscr	83.44	82.70	84.13	0.67	83.44	82.70	84.13	0.67	82.66	86.89	78.69	0.66
HHDensemble	84.05	79.32	88.49	0.68	82.41	79.75	84.92	0.65	85.25	85.25	85.25	0.70

C2.1 PELVIS - TABLE C2-1

Of the 5 bone models, Pelvic models were the least accurate. Based on *k*-fold CV results, the best Pelvic model is PelvisKnn. However, on inspection of the training accuracy estimate, it becomes clear that this model has been overfit. The PelvisEnsemble, PelvisTree and PelvisNb models are all more accurate than the PelvisDiscr model, but they are also all present with large sex biases (~10-12%). Thus, the PelvisDiscr model is selected as the best fit Pelvic model.

The optimisation function used in model training of PelvisTree has grown a tree excluding TAD as a node (Figure C2.1c) whereas some of the trees in the PelvisEnsemble model (Figure C2.1b) have grown a second node and this has likely contributed to the slight increase in model accuracy from PelvisTree to PelvisEnsemble. The DFA model (Figure C2.1d) has used both variables, but AD has a larger contribution than TAD to the model's success. *K*-fold CV accuracy values for PelvisDiscr are higher than univariate ADdiscr and TADdiscr.

When the incorrect predictions are superimposed on the test data in Figure C2.1a. It is lucid that PelvisDIscr, the best fit pelvic model, struggles to correctly classify sex where there is overlap between males and females. Male test observations are more variable and have a wider range of values for both predictor variables than female observations

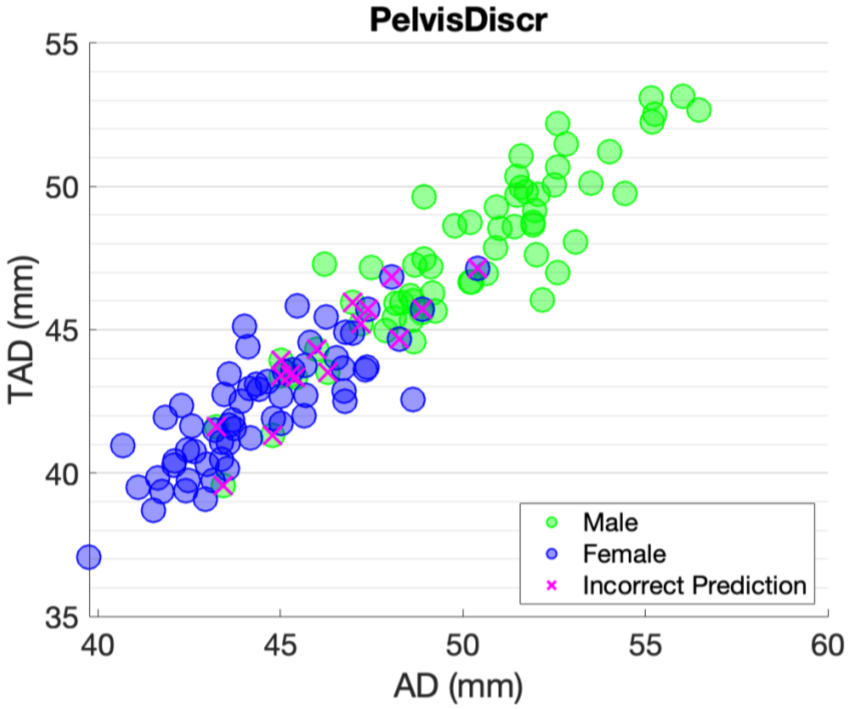


Figure C2.1a Incorrect holdout predictions plotted over the holdout subset to indicate areas where the model fails.

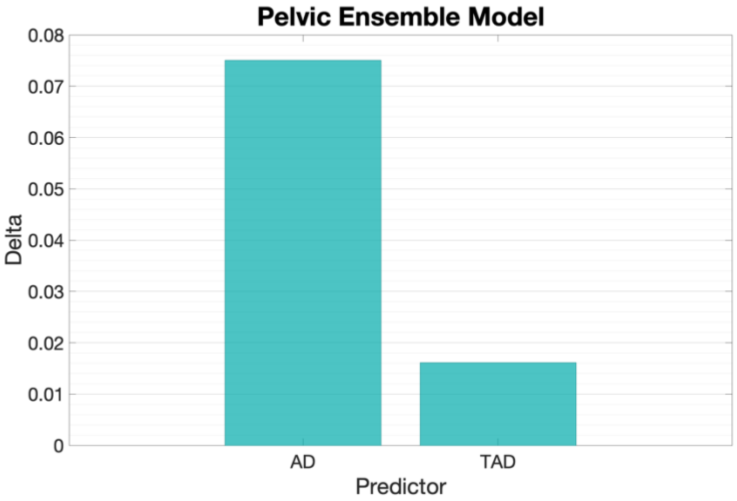


Figure C2.1b Contribution of different predictors to the PelvisEnsemble model.

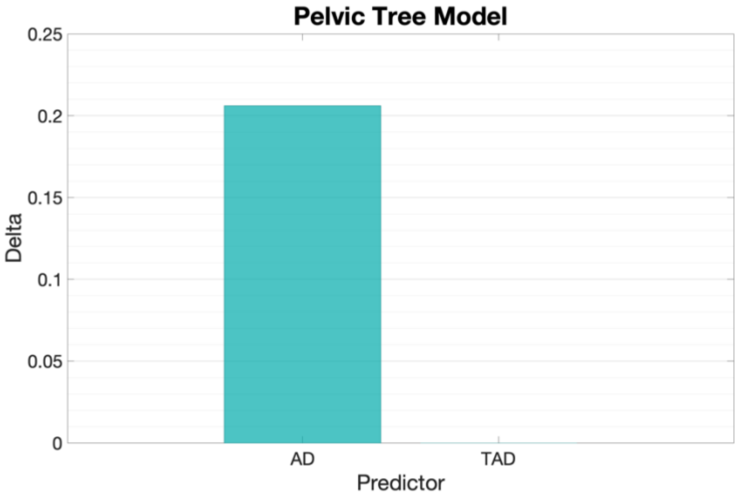


Figure C2.1c Contribution of different predictors to the PelvisTree model.

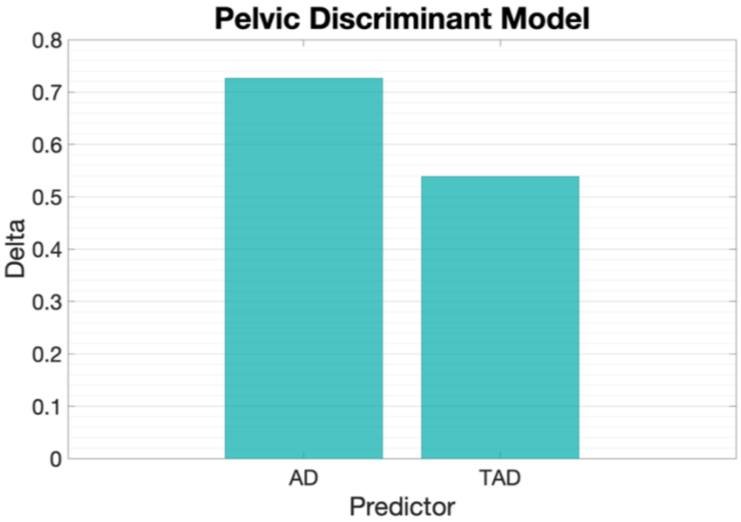


Figure C2.1d Contribution of different predictors to PelvisDiscr model.

C2.2 FEMUR - TABLE C2-2

K-fold CV accuracies for FemurNb and FemurTree models are lower than when FHD or THD are used alone. It is no surprise that the FemurTree model is no more successful than a univariate model as it only includes a root node (Figure C2.2a) and has excluded the other predictors. The sectioning point and, consequently, training accuracies for the FemurTree and THDtree models are identical.

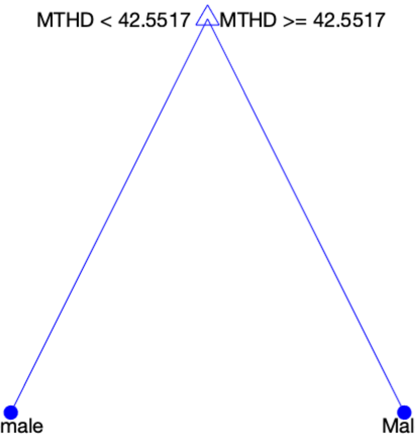


Figure C2.2a Visualisation of PelvisTree model (MTHD = mean transverse femoral head diameter).

PelvisDiscr is the most successful pelvic model as a result of its high accuracy estimates. All 5 predictors in the FemurDiscr model contribute to model success (Figure C2.2b) with BB contributing the most and MCL contributing the least. FemurDiscr *k*-fold CV accuracy is higher than for any univariate femoral model.

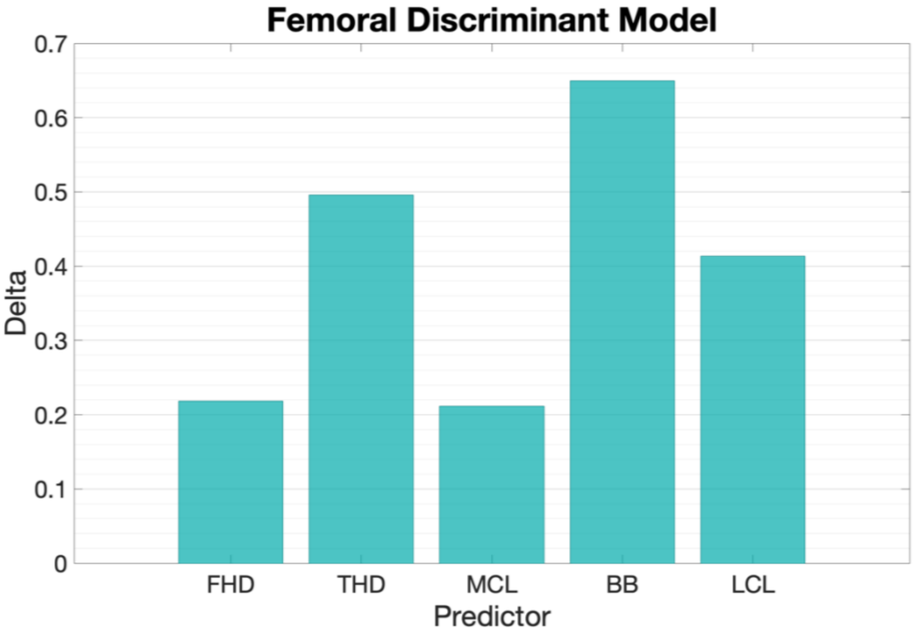


Figure C2.2b Contribution of predictors to the multivariate FemurDiscr model.

C2.3 TIBIA - TABLE C1-8

See univariate TPB in Appendix C1.8

C2.4 SCAPULA - TABLE C2-4

Bivariate scapula models were more accurate than all univariate GB models and either equal to or better than univariate GL models. The training accuracies for GLtree and ScapulaTree are the same because the ScapulaTree model has optimised to only use GL and has the same sectioning point as GLtree. The most accurate bone model, superior to all univariate models and multivariate bone models, is the scapula discriminant model which has a *k*-fold CV accuracy of 90.3%. Both GL and GB contribute to the success of the model as shown in Figure C2.4 although, GL is the larger contributor.

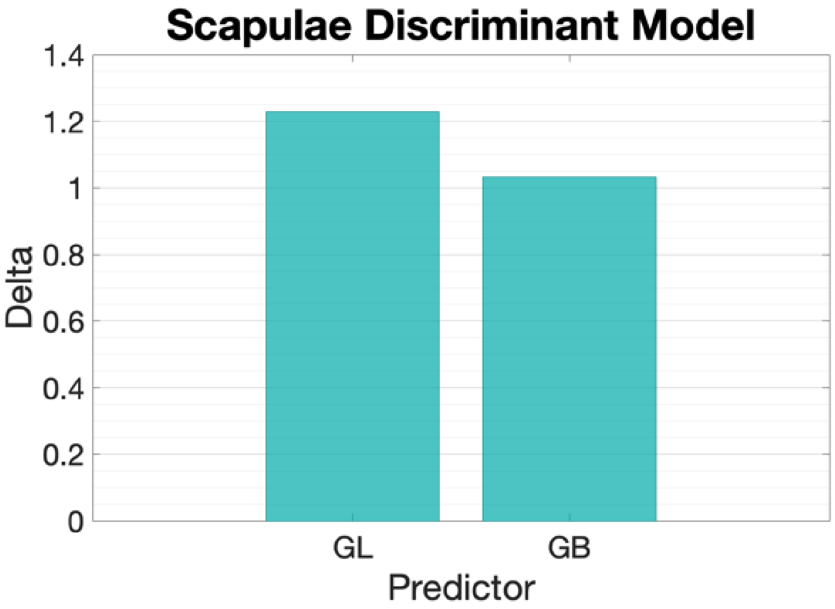


Figure C2.4 Contribution of descriptor variables to ScapulaDiscr model.

Table C2-1 Goodness of fit for all multivariate pelvic mensuration models. Pelvic bone model includes AD and TAD.

	Training Accuracy				***K*-fold Accuracy**				**Holdout Accuracy**			
Model	All	Male	Female	Kappa	All	Male	Female	Kappa	All	Male	Female	Kappa
PelvisNb	84.78	79.12	90.27	0.70	84.78	79.12	90.27	0.70	87.79	90.16	85.48	0.76
PelvisKnn	100.00	100.00	100.00	1.00	85.38	82.33	88.33	0.71	87.77	91.80	83.87	0.76
PelvisTree	85.57	79.92	91.05	0.71	84.98	79.92	89.88	0.70	84.55	85.25	83.87	0.69
PelvisDiscr	84.58	84.74	84.44	0.69	84.58	83.94	85.21	0.69	86.96	91.80	82.26	0.74
PelvisEnsemble	85.97	81.12	90.66	0.72	85.18	79.92	90.27	0.70	83.73	85.25	82.26	0.67

Table C2-2 Goodness of fit for all multivariate femoral mensuration models. Femoral bone model includes FHD. THD. MCL. BB and LCL.

	Training Accuracy				***K*-fold Accuracy**				**Holdout Accuracy**			
Model	All	Male	Female	Kappa	All	Male	Female	Kappa	All	Male	Female	Kappa
FemurNb	85.25	83.26	87.15	0.70	85.25	83.26	87.15	0.70	78.11	81.36	75.00	0.56
FemurKnn	86.27	84.94	87.55	0.73	86.27	84.94	87.55	0.73	83.13	88.14	78.33	0.66
FemurTree	87.32	83.68	90.76	0.75	84.43	84.52	84.34	0.69	83.15	86.44	80.00	0.66
FemurDiscr	87.70	86.61	88.76	0.75	87.50	86.19	88.76	0.75	83.96	89.83	78.33	0.68
FemurEnsemble	88.32	87.45	89.16	0.77	85.86	85.77	85.94	0.72	82.28	88.14	76.67	0.65

Table C2-4 Goodness of fit for all multivariate scapulae mensuration models. Scapula bone model includes GL and GB.

	Training Accuracy				***K*-fold Accuracy**				**Holdout Accuracy**			
Model	All	Male	Female	Kappa	All	Male	Female	Kappa	All	Male	Female	Kappa
ScapulaNb	90.08	89.34	90.80	0.80	89.47	88.11	90.80	0.79	85.82	86.67	85.00	0.72
ScapulaKnn	89.88	88.93	90.80	0.80	90.08	88.93	91.20	0.80	86.67	86.67	86.67	0.73
ScapulaTree	89.88	92.21	87.60	0.80	88.66	90.98	86.40	0.77	84.14	86.67	81.67	0.68
ScapulaDiscr	90.49	90.16	90.80	0.81	90.28	90.16	90.40	0.81	85.82	86.67	85.00	0.72
ScapulaEnsemble	90.08	88.52	91.60	0.80	87.85	86.48	89.20	0.76	85.00	85.00	85.00	0.70

C2.5 HUMERUS - TABLE C2-5

Bivariate humeral models (using EB and HHD as predictors) were superior to all univariate humeral predictors with accuracies and kappa values being higher for all algorithms compared to univariate EB and HHD models. Training and *k*-fold CV sex bias for all humeral models is low. All humeral models are similar in their *k*-fold CV accuracies (<0.6% difference) and are good predictors of sex with all *k*-fold CV accuracies higher than 89%. The humeral model selected as the best fit is HumerusDiscr as it has the most consistently high accuracies and kappa statistics across training, *k*-fold, and Ho cross-validation.

C3 JOINT MODELS

Multivariate joint-specific models were trained and optimised for the Hip, Knee and Shoulder. Results pertaining to best fit models are set out in the sections to follow along with comparisons to univariate models and bone-specific models.

C3.1 HIP - TABLE C3-1

The best fit multivariate hip model is the HipDiscr model. This model is a quadratic discriminant model. It has the highest *k*-fold CV kappa (0.73) of all Hip models and minimal sex bias. HipDiscr (86.5%) outperformed univariate pelvic models, ADdiscr (84.4%) and TADdiscr (82.8%), but not femoral univariate models, FHDdiscr (86.9%) and THDdiscr (86.7%) although, the differences are very small and likely due to differing datasets. When compared to the bivariate Pelvic model, *k*-fold CV accuracy of HipDiscr is higher than for PelvisDiscr (84.6%). The inclusion of 2 extra predictor variables was thus able to improve prediction accuracy by 2.0%.

C3.2 KNEE - TABLE C3-2

All Knee models have similar *k*-fold CV accuracies. The Knee KNN, tree and ensemble models share a CV kappa statistic of 0.77. The KNN model has a perfect training accuracy and is thus likely overfit. For KneeTree, the final model only uses TPB. TPBtree and KneeTree share a sectioning point of 71.52mm but KneeTree is slightly more accurate. TPBtree was trained using a sample of 452 compared to 442 for KneeTree. This result emphasises the how easily the accuracy of binary decision trees is influenced by their training dataset.

The best fitting Knee model is KneeDiscr. The KneeDiscr model has the highest kappa statistic of all Knee models, consistent results across the 3 cross-validation types and minimal sex bias. The multivariate KneeDiscr model (trained using MCL, BB, LCL and TPB) is a quadradic discriminant model and performs better than univariate MCL, BB, LCL and TPB discriminant models.

C3.3 SHOULDER - TABLE C3-3

The least effective shoulder model under *k*-fold CV is the tree model with 86.7% accuracy. The optimised ShoulderTree model grew to only include GL, with a sectioning point of 36.755mm (identical to the GLtree model). The Shoulder model with the highest *k*-fold CV accuracy is the NB model but it has moderate sex bias. With 0.4% lower *k*-fold CV accuracy but no sex bias, the quadratic ShoulderDiscr model is chosen as the best fit shoulder model. Figure C3.3 shows that the 3-predictor ShoulderDiscr model struggles to accurately predict the sex of individuals in the zone of overlap between the sexes.

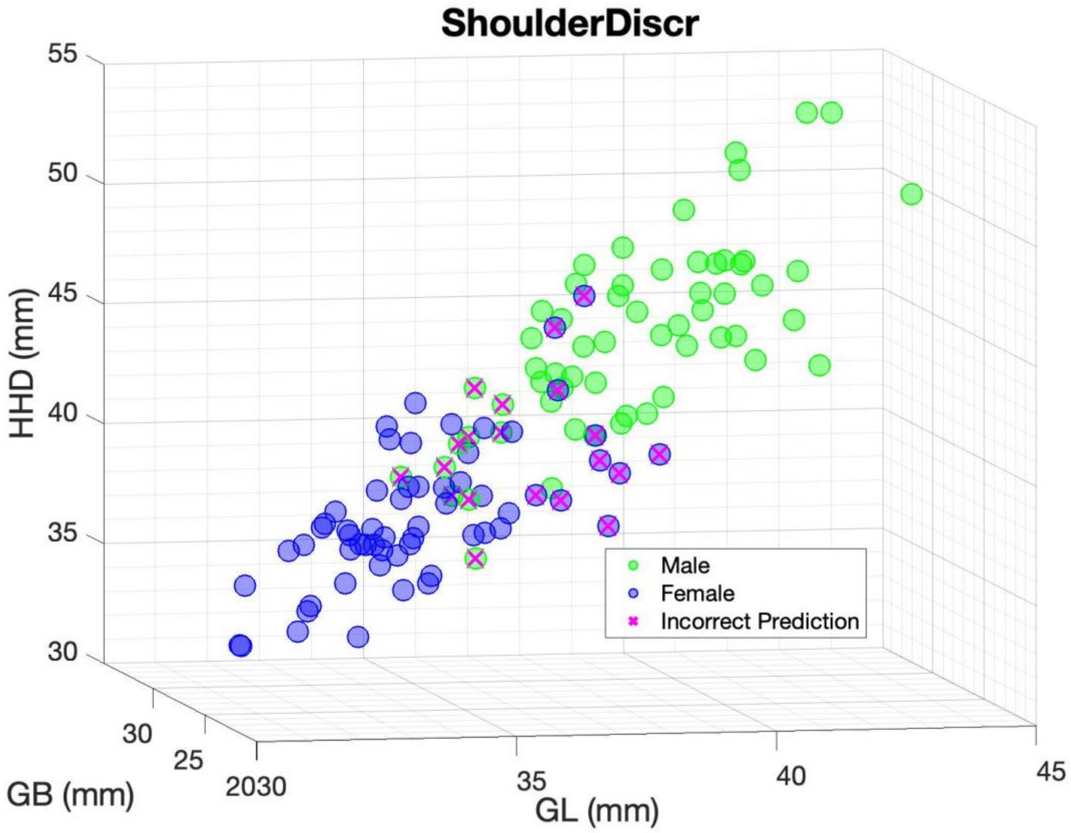


Figure C3.3 Incorrect predictions using the ShoulderDiscr model superimposed over the training dataset.

Table C2-5 Goodness of fit for all multivariate humeral mensuration models. Humeral bone model includes EB and HHD

	Training Accuracy				***K*-fold Accuracy**				**Holdout Accuracy**			
Model	All	Male	Female	Kappa	All	Male	Female	Kappa	All	Male	Female	Kappa
HumerusNb	90.80	90.30	91.27	0.82	89.57	88.61	90.48	0.79	86.78	90.16	83.61	0.74
HumerusKnn	89.98	89.03	90.87	0.80	89.57	88.61	90.48	0.79	85.99	88.52	83.61	0.72
HumerusTree	90.39	89.45	91.27	0.81	89.57	88.19	90.87	0.79	85.20	86.89	83.61	0.70
HumerusDiscr	90.18	90.72	89.68	0.80	89.98	90.30	89.68	0.80	88.37	93.44	83.61	0.77
HumerusEnsemble	91.62	91.98	91.27	0.83	89.37	90.72	88.10	0.79	87.58	91.80	83.61	0.75

Table C3-1 Goodness of fit for all multivariate hip joint mensuration models. Hip joint model includes AD. TAD. FHD and THD.

	Training Accuracy				***K*-fold Accuracy**				**Holdout Accuracy**			
Model	All	Male	Female	Kappa	All	Male	Female	Kappa	All	Male	Female	Kappa
HipNb	85.92	85.11	86.69	0.72	86.13	85.53	86.69	0.72	83.84	89.66	78.33	0.68
HipKnn	100.00	100.00	100.00	1.00	86.34	84.26	88.31	0.73	84.70	89.66	80.00	0.70
HipTree	87.37	83.83	90.73	0.75	85.30	82.98	87.50	0.71	83.86	87.93	80.00	0.68
HipDiscr	86.54	85.96	87.10	0.73	86.54	85.96	87.10	0.73	83.84	89.66	78.33	0.68
HipEnsemble	87.58	84.26	90.73	0.75	86.13	84.68	87.50	0.72	83.86	87.93	80.00	0.68

Table C3-2 Goodness of fit for all multivariate knee joint mensuration models. Knee joint model includes MCL. BB. LCL and TPB

	Training Accuracy				***K*-fold Accuracy**				**Holdout Accuracy**			
Model	All	Male	Female	Kappa	All	Male	Female	Kappa	All	Male	Female	Kappa
KneeNb	85.75	83.81	87.50	0.71	85.97	83.81	87.93	0.72	80.95	79.63	82.14	0.62
KneeKnn	100.00	100.00	100.00	1.00	88.69	87.14	90.09	0.77	86.17	92.59	80.36	0.73
KneeTree	89.14	92.38	86.21	0.78	88.24	91.90	84.91	0.77	86.99	96.30	78.57	0.75
KneeDiscr	89.59	89.05	90.09	0.79	89.37	88.10	90.52	0.79	87.05	94.44	80.36	0.75
KneeEnsemble	93.67	92.86	94.40	0.87	88.69	89.52	87.93	0.77	86.17	92.59	80.36	0.73

C4 ALL PREDICTORS - TABLE C4

'All-predictors' (AP) models were trained using all 12 predictor variables. The worst performing AP model is the tree model, APtree. Despite being trained using 12 predictors, APtree grew to have only 2 nodes, GL and TPB, as shown in Figure C4a. These 2 predictors were selected by the MRMR algorithm (Figure C4a) as the 2 'most important' predictors. In addition to being the worst performing AP model, APtree was outperformed by the univariate GLtree model.

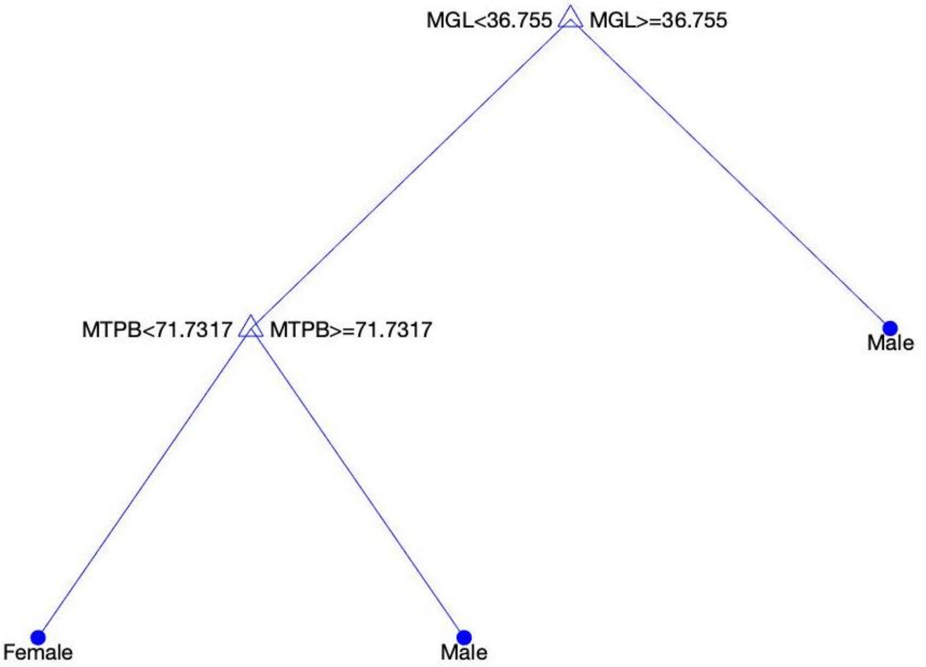


Figure C4a All predictors classification tree model visualisation.

Different ensemble AP models were trained. The 2 different types of ensemble tree outperformed the APtree model and minimised the sex classification bias which was present in the APtree model. Although ensemble tree models were able to outperform a standard decision tree, the far less complex APlinearDiscr model was superior to both ensemble trees. The APsubspaceEnsembleDiscr model was equally as accurate as the standard linear discriminant model but, given its added complexity, inferior to APlinearDiscr. The fourth ensemble model, APsubspaceKNN, fell prey to overfitting and was thus less successful that the standard KNN model.

The best performing AP model is the APminkowskiKNN model which has a *k*-fold CV accuracy of 93.0% and no sex bias. When considering model complexity and ease of use however, the

KNN model is not worth the added 0.2% accuracy which it affords over the LDA model. The APlinearDiscr model with 92.8% *k*-fold CV accuracy is thus selected as the best fit AP model. The main contributors to this model are GL and TPB, as seen in Figure C4b with FHD and BB having the smallest contributions.

Of the best fit models selected for univariate and multivariate predictor combinations, APlinearDiscr is 2.19% more accurate than the next most accurate model overall which is ScapulaDiscr (GL + GB).

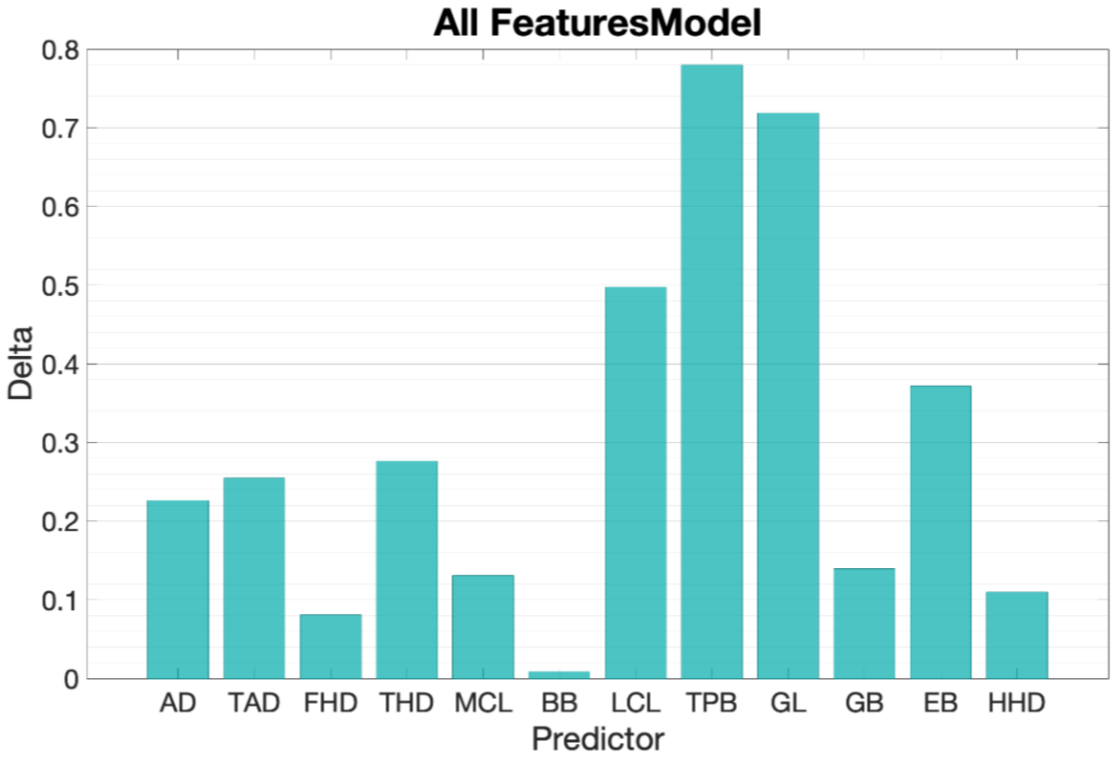


Figure C4b Relative predictor contributions to APlinearDiscr model.

Table C3-3 Goodness of fit for all multivariate shoulder joint mensuration models. Shoulder joint model includes GL. GB and HHD.

	Training Accuracy				***K*-fold Accuracy**				**Holdout Accuracy**			
Model	All	Male	Female	Kappa	All	Male	Female	Kappa	All	Male	Female	Kappa
ShoulderNb	89.79	87.12	92.31	0.80	89.79	87.12	92.31	0.80	87.43	90.00	85.00	0.75
ShoulderKnn	100.00	100.00	100.00	1.00	89.17	85.84	92.31	0.78	87.48	88.33	86.67	0.75
ShoulderTree	89.79	92.27	87.45	0.80	86.67	87.55	85.83	0.73	84.09	86.67	81.67	0.68
ShoulderDiscr	89.38	88.84	89.88	0.79	89.38	88.84	89.88	0.79	83.33	83.33	83.33	0.67
ShoulderEnsemble	89.79	90.99	88.66	0.80	88.96	90.13	87.85	0.78	83.33	83.33	83.33	0.67

Table C4 Goodness of fit for AP (all predictor) models.

	Training Accuracy				***K*-fold Accuracy**				**Holdout Accuracy**			
Model	All	Male	Female	Kappa	All	Male	Female	Kappa	All	Male	Female	Kappa
APbaggedEnsembleTree	96.14	96.37	95.95	0.92	90.84	89.64	91.89	0.82	90.30	92.16	88.68	0.81
APboostedEnsembleTree	97.59	96.37	98.65	0.95	91.81	91.19	92.34	0.84	88.38	90.20	86.79	0.77
APsubspaceEnsembleDiscr	93.01	93.78	92.34	0.86	92.77	93.78	91.89	0.86	87.37	90.20	84.91	0.75
APsubspaceEnsembleKnn	100.00	100.00	100.00	1.00	91.81	91.71	91.89	0.84	87.37	90.20	84.91	0.75
APnb	91.08	91.19	90.99	0.82	89.88	89.12	90.54	0.80	84.44	88.24	81.13	0.69
APminkowskiKNN	94.94	93.78	95.95	0.90	93.01	92.23	93.69	0.86	88.38	90.20	86.79	0.77
APlinearDiscr	93.49	93.26	93.69	0.87	92.77	92.75	92.79	0.85	85.35	90.20	81.13	0.71
APtree	91.81	89.64	93.69	0.84	88.43	82.38	93.69	0.77	85.74	82.35	88.68	0.71

APPENDIX D: OPTIMAL NUMBER OF PREDICTORS

D1 TREE MODEL

For tree models, at lower sample sizes, there is no clear relationship between error and number of predictors. At the maximum sample size (350), the univariate mode is least accurate whilst the accuracy for models with more than 4 predictors are all almost the same. Given the relative improvement in accuracy with any given number of predictors, a 3-predictor tree model is likely sufficient to maximise accuracy and minimise redundancy.

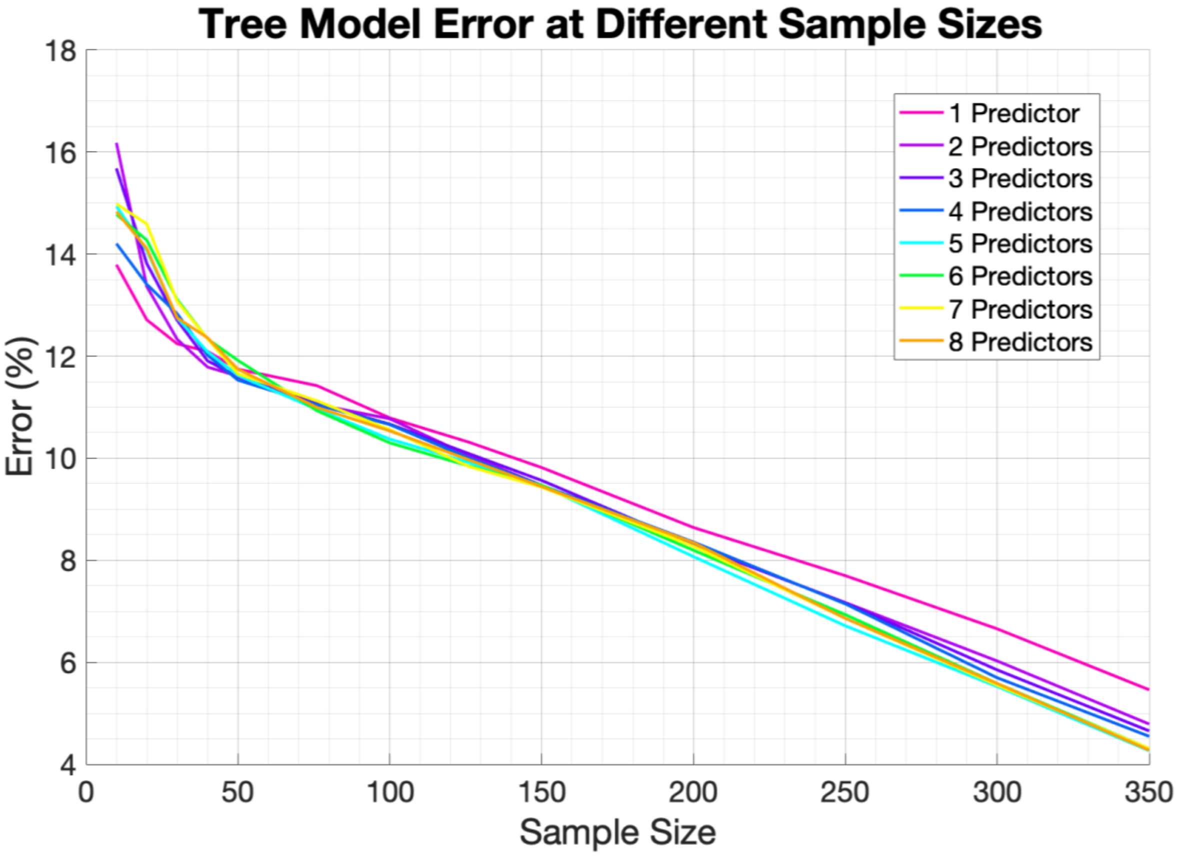


Figure D1 Mean model error for decision tree models trained 1 000 times with 1 predictor (GL), 2 predictors (+TPB), 3 predictors (+ LCL), 4 predictors (+ EB), 5 predictors (+FHD), 6 predictors (+AD), 7 predictors (+ THD) and 8 predictors (+ GB) at increasingly large sample sizes.

D2 ENSEMBLE MODEL

The curve for bagged ensemble tree models is quite different from the decision tree curve in shape. The number of predictors in the model has a far more pronounced impact on ensemble model error compared to tree model error. Increasing the number of predictors from 1 to 2 has a marked impact on the model error at all sample sizes. The difference between 2 and 3 predictors is still notable but much smaller. The optimal number of predictors for a bagged ensemble forest is likely 5. Error does not decline much when the number of predictors exceeds 5.

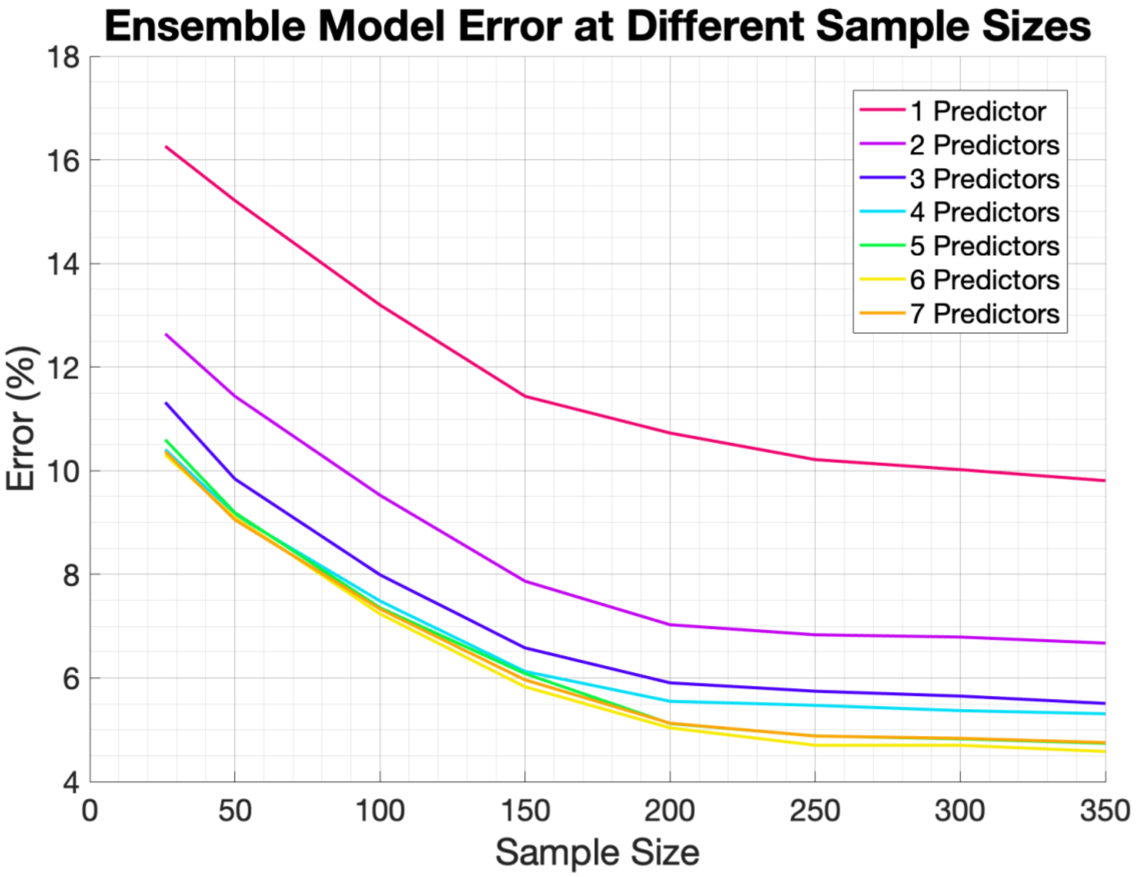


Figure D2 Mean model error for bagged ensemble tree models trained 1 000 times with 1 predictor (GL), 2 predictors (+TPB), 3 predictors (+ LCL), 4 predictors (+ EB), 5 predictors (+FHD), 6 predictors (+AD), 7 predictors (+ THD) and 8 predictors (+ GB) at increasingly large sample sizes.

D3 NAÏVE BAYES MODEL

When NB models are trained with an increasing number of predictors, the curves do not follow a consistent trend. With tree models and ensemble models, as the number of predictors increased, model error decreased. This is not true for NB models. At a sample size of between 250 and 350 observations, a 5-predictor model seems to have the least error whilst a 2- or 3-predictor model has more error than a univariate model. The reason for this trend is unclear however, it does form part of the justification for why NB models were not selected as best-fit models.

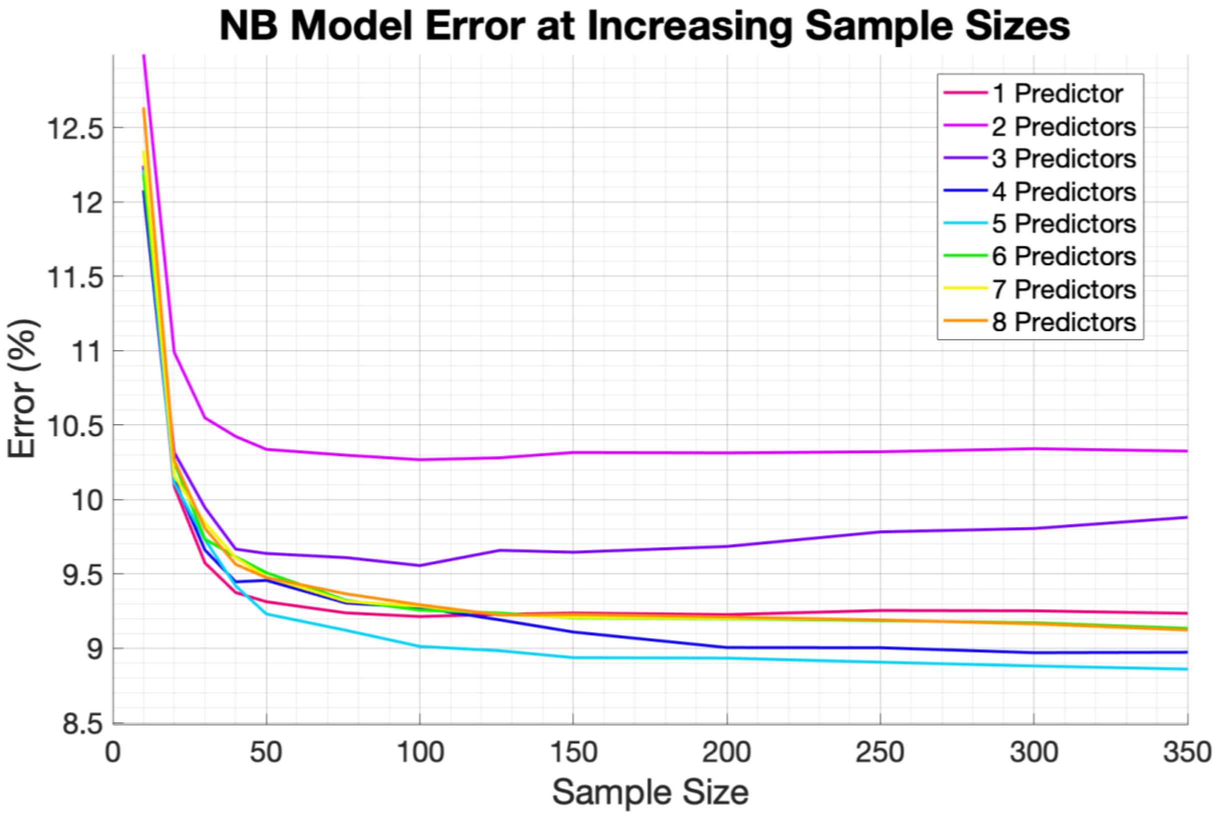


Figure D3 Mean model error for Naïve Bayes models trained 1 000 times with 1 predictor (GL), 2 predictors (+TPB), 3 predictors (+ LCL), 4 predictors (+ EB), 5 predictors (+FHD), 6 predictors (+AD), 7 predictors (+ THD) and 8 predictors (+ GB) at increasingly large sample sizes.

D4 KNN MODEL

When KNN models are trained with an increasing number of predictors, there is a direct relationship between number of predictors and model error. As the number of predictors increase, the model error decreases. The differences are most pronounced with 1, 2 and 3-predictor models and less so with more than 3-predictors. A 3-predictor KNN model is likely optimal for maximal accuracy and minimal redundancy.

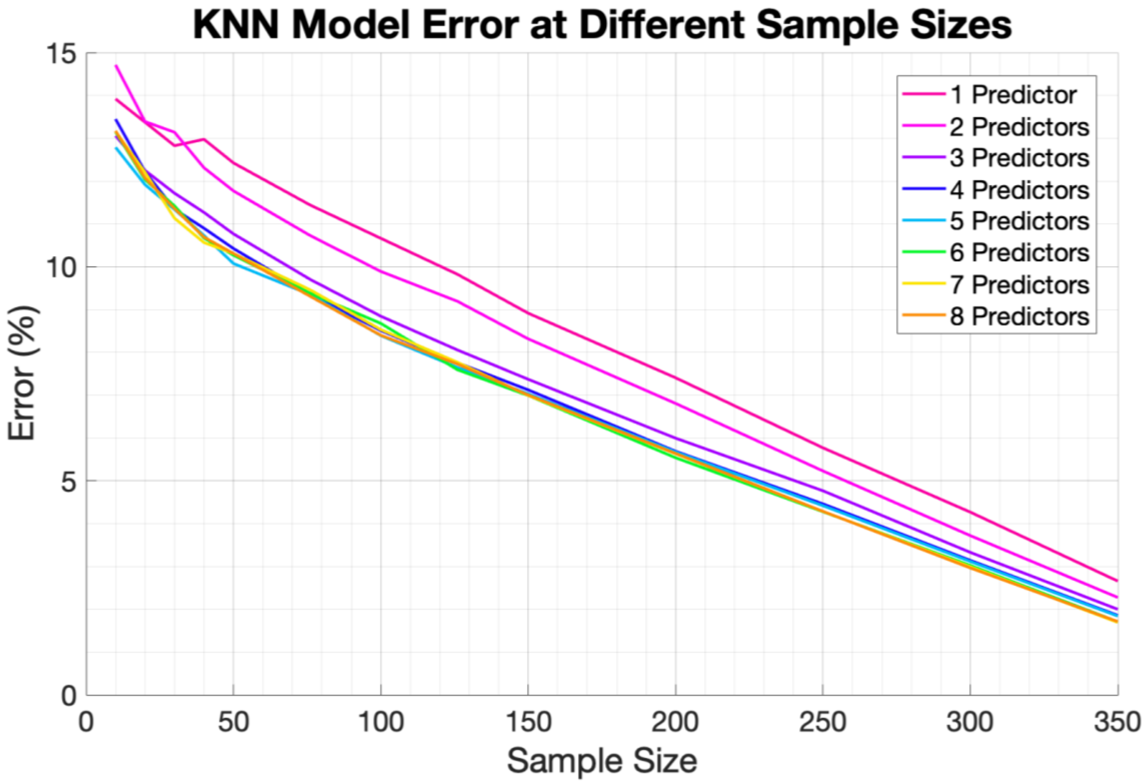


Figure D4 Mean model error for KNN models trained 1 000 times with 1 predictor (GL), 2 predictors (+TPB), 3 predictors (+ LCL), 4 predictors (+ EB), 5 predictors (+FHD), 6 predictors (+AD), 7 predictors (+ THD) and 8 predictors (+ GB) at increasingly large sample sizes.