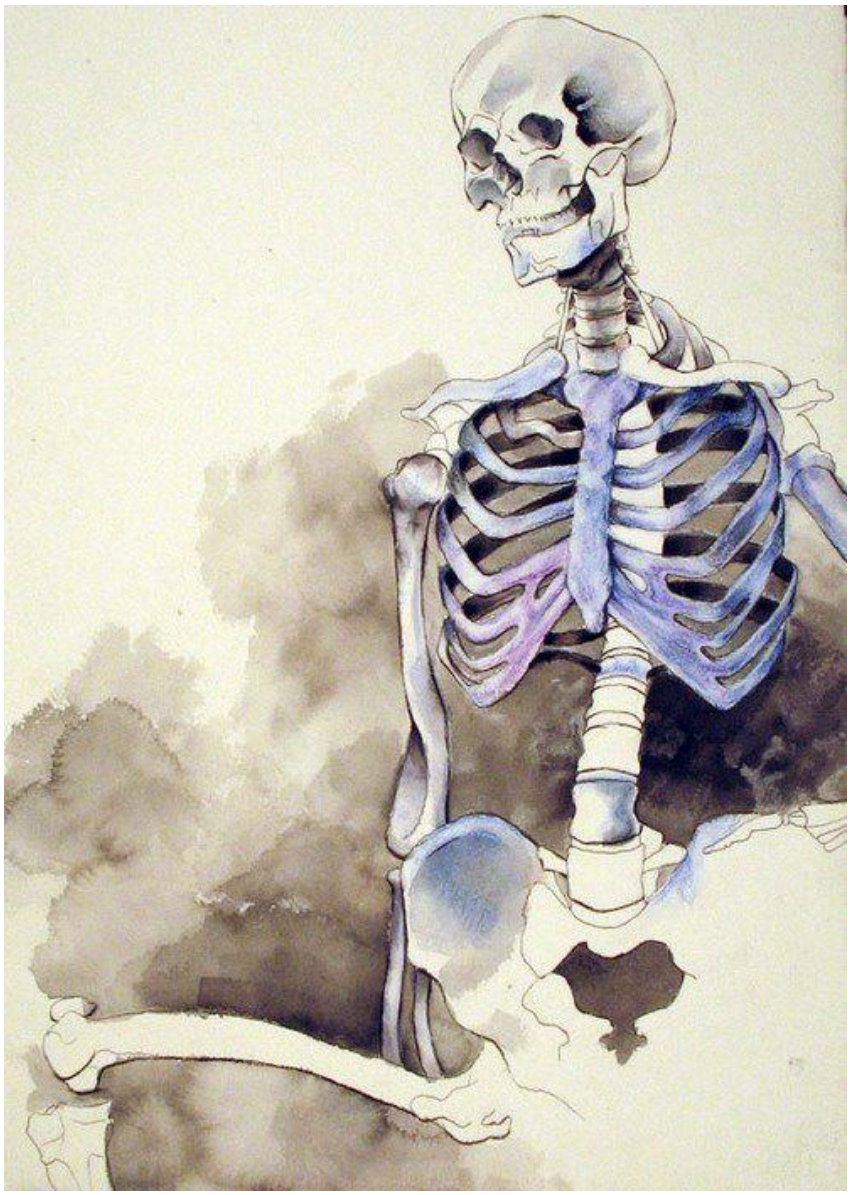
Age-at-death estimation through neural networks: an evaluation of DRNNAGE software for age-at-death estimation on a Dutch medieval skeletal sample

Reus, Babette

AGE-AT-DEATH ESTIMATION THROUGH NEURAL NETWORKS:

An evaluation of DRNNAGE software for age-at-death estimation on a Dutch medieval skeletal sample



Babette Reus

Leiden University | Dr. Schrader

Front title image is AI-generated with Midjourney

Age-at-death estimation through neural networks:

an evaluation of DRNNAGE software for age-at-death estimation on a Dutch medieval skeletal sample

Babette Reus
s3378810
Master Thesis Archaeological Science (1084VTSY)
Dr. Schrader (Supervisor)
Leiden University, Faculty of Archaeology
Leiden, 29-12-2023, Final

Universiteit Leiden
The Netherlands

Table of Contents

Table of Figures

Table of Tables

Acknowledgements

I want to thank my supervisor Dr. Schrader, for the support, time, and guidance. I want to thank the Laboratory for Human Osteoarchaeology (Leiden University) in Leiden for allowing me to work with the MB11 skeletal collection and providing me with the archival documents. I express my gratitude to all contributors who are not mentioned that played a role in the completion of this thesis.

Introduction

Traditional skeletal age-at-death estimation concerns the evaluation of morphological changes in the skeleton. Age-at-death estimation for subadults relies on systematic and constant changes in skeletal morphology during development, providing prominent age-related indicators (White & Folkens, 2005, p.365). In adults, age-at-death estimation is challenging because the chronological time, measured in years, often deviates from biological age, which reflects the body's state in terms of development and degeneration (Boldsen *et al.,* 2002, p.73; Blau & Ubelaker, 2016, p.273). Osteologists and forensic anthropologists estimate the *biological age* based on morphological changes, which are susceptible to confounding factors such as sex, ancestry, and individual variation (Ubelaker & Khosrowshahi, 2019, p.2). However, this approach introduces bias and presents challenges in estimating the ages of older individuals, especially those over 50 years (Boldsen *et al.,* 2002, p.73; Buk *et al.,* 2012, p.1; Obertova *et al.,* 2020, p.209). The debate on how these methods should interact to form a precise age estimate is still ongoing, and adult age estimation, remains a complex task with a potential error of approximately 12 years for adults (Navega *et al.,* 2022, p.2).

In addition, the estimation of age may be sensitive to interobserver errors, which occur when multiple researchers assess an individual's age and reach differing conclusions (Obertova *et al.,* 2020, p.209). Traditional age-at-death estimation methods often require the support of other methods to provide accurate outcomes. One challenge involves age estimates mirroring the skeletal age trajectory and characteristics of known age reference individuals used to create these estimation methods. This phenomenon is referred to as “age mimicry” and a proposed solution involves transition analysis which designates the morphological changes in the skeleton throughout the lifespan into several stages (Schanandore *et al.,* 2021, p.65; Rizos *et al.,* 2023, p.2). Transition analysis examines how skeletal features change considering the dynamic progression of skeletal features which is important because skeletal features are not uniform among individuals. On the other hand, it is suggested that the performance of transition analysis is equal compared to traditional methods and is therefore not superior (Rizos e*t al.,* 2023, p.2)

The multifactorial method is suggested to aid in this problem, which involves considering and combining multiple factors or indicators, such as various skeletal features or multiple anatomical regions, to develop a more comprehensive and accurate assessment of age-at-death (Bedford *et al.,* 1993, p.287). This approach recognizes the complexity of aging processes and aims to improve estimation accuracy by combining a diverse set of contributing factors rather than relying on a single indicator. The multifactorial method distinguishes itself from traditional approaches by aiming to encompass numerous anatomical regions and features. This approach aims to acquire an accurate estimate by considering multiple age trajectories across different anatomical regions, in contrast to concentrating on the age trajectory of a single region such as the pubic symphysis method by Suchey and Brooks (Ch 2.1.1).

In response to the growing need for an accurate age estimation method, Navega *et al.* (2022) have developed a multifactorial approach for precise age-at-death estimation. What is interesting is that they implemented Artificial Intelligence (AI) to compute the estimates, in this case, a randomized deep Neural Network (NN) was utilized. A NN is a computational model created to process

information through interconnected neurons, similar to the human brain, allowing it to learn and make predictions from data.

The NN that Navega *et al.* (2022) developed and implemented in their study was used to estimate age-at-death in a Portuguese sample of 500 individuals of known age, where they combined several skeletal features to accurately predict age-at-death (p.1). They are very supportive of their results, in which they claim that the NN can estimate age-at-death with a mean absolute error (MAE) of approximately six years throughout the lifespan for both young (adult) and old individuals. As the impact of AI has become increasingly evident in our daily lives, its application in osteology is fascinating. Because this method claiming high accuracy, it is useful to investigate its performance across diverse population contexts and assess its replicability. The Middenbeemster collection (MB11), a Dutch medieval archaeological population, was used to test the validity of this method. The inclusion of archival documents with recorded age-at-death information enhances the advantage of utilizing this collection for validating novel age-at-death estimation methods.
Navega *et al.*, (2022) have developed the DRNNAGE (Deep Randomized Neural Network for Age) software model, which allows implementation of this method for other studies. The application of NNs in age-at-death estimation requires expertise in multiple contexts such as statistics, programming, and osteology. Therefore, with emphasis on age-at-death estimation, it is necessary to review whether the implementation of NNs is worthwhile in the field of osteology. Overall, this thesis aims to investigate the potential of deep random NNs for age-at-death estimation in adult skeletal remains. The results of this study have implications for forensic anthropology and bio archaeology, contributing to the development of more accurate and reliable methods for age-at-death estimation.

1.1 Research aim

Recent studies (Corsini *et al*., 2004; Buk *et al.*, 2012; Czibula *et al*., 2016; Toneva *et al.*, 2017; Navega *et al.*, 2018, 2022) have demonstrated that AI implementations such as NNs, can improve the accuracy and precision of age-at-death estimation in adult skeletal remains. Age-at-death assessment is a crucial step in the identification of skeletal remains, yet many traditional methods fall short in providing precise estimates, often resulting in broad age ranges. As the challenge of aligning chronological and biological age remains, it is becoming increasingly difficult to estimate elderly individuals (Buk *et al.*, 2012, p.1, Navega *et al.*, 2022, p.2). In pursuit to gain understanding of past lives, numerous age estimation methods have evolved over time. Osteologists and forensic anthropologists play a crucial role in this journey, contributing to the attribution of crucial information, such as stature, sex, pathology, and age at death. However, even with existing methods, age-at-death remains an estimate. In short, the evaluation of the application of NNs in age-at-death estimation and its adaptability to diverse populations and contexts will be examined, driven by the immediate need for a method that addresses to all current problems

This thesis attempts to investigate the use of (deep random) NNs in age-at-death estimation of adult skeletal remains. The methodology involves the collection of morphological and degenerative traits from archaeological skeletal remains to create a dataset, a data collection strategy, data pre-processing, model training, and statistical analysis. The expected results of this study are improved accuracy and consistency of age-at-death estimation over traditional methods, relevant to an

archaeological Dutch medieval population. The results of this study have implications for the future of AI in forensic anthropology and osteology.

1.2 Research questions

This thesis aims to determine whether the trained deep random NN (DRNNAGE software) can accurately estimate the age-at-death of adult skeletal remains from a Dutch medieval sample.

The following sub questions were addressed:

- Which factors can influence age-at-death estimation using deep random NNs?
- How does the accuracy of trained deep random NNs compare with the traditional methods of age-at-death estimation in adult skeletal remains?
- How can trained deep random NNs be used to improve age-at-death estimation in diverse populations and contexts?

1.3 Thesis structure

This thesis comprises six chapters, each of which will present a different portion of the study. *Chapter 1* is divided into four sections and presents a general introduction to the topic of age-at-death estimation and the current problems it entails. *Section 1.1* and *1.2* will address the research aim and the research questions of this study. The chapter ends with the thesis structure in *section 1.3. Chapter 2* is divided into four sections to provide a deeper understanding of the background necessary for this study. *Section 2.1,* is an introduction about the traditional age estimation methods that exist, which are discussed in more detail in *section 2.2*. In *section 2.3*, a conclusion regarding the traditional age-at-death estimation is drawn. *Section 2.4* will end the chapter and provides more background on NNs. *Chapter 3* is divided into six sections that present the methodology. *Section 3.1* addresses feature selection, section 3.2 will cover the methodological strategy, while *section 3.4* and *3.5* cover data pre-processing and data analysis, respectively. This chapter ends with the conclusion in *section 3.6*. *Chapter 4* is divided into four sections, which present all the results obtained during the study. *Section 4.1,* presents the estimated DRNNAGE age-at-death against the archival age, *section 4.2* presents the results obtained from comparing the estimated DRNNAGE age-at-death difference between females and males, *section 4.3* presents the results obtained from comparing the estimated DRNNAGE age-at-death between age groups, and *section 4.4* presents the obtained results from evaluating the estimated DRNNAGE age-at-death against the preservation score. *Section 4.5* will conclude the paper. *Chapter 5* is divided into five sections that discuss the collected results and provide critical interpretations. *Section 5.1* displays a summary of the main findings and *section 5.2* will provide interpretations of the data and aims to address the research questions from *section 1.2* in *section 5.3*. *Section 5.4* covers the limitations of the study. *Chapter 6* completes the thesis as it draws final conclusions about the study and addresses future research.

Background

The first section of this chapter will present traditional methods for estimating the age-at-death. Subsequently, the background of NNs is discussed, highlighting its potential applications in the field of osteology and forensic anthropology. Following that, studies that have utilized AI methods are explored by discussing their findings and implications.

2.1 Traditional methods for age-at-death estimation

From the 16th century onwards, numerous methods to estimate age-at-death have been created, most of which focus on a specific region of the skeleton (Meindl & Lovejoy, 1985, p.57; White & Folkens, 2005, p.369). The aim is to develop a precise method that delivers accurate results without requiring a full-skeletal examination. However, multiple studies share the concern that there is currently no method to meet this demand (Lovejoy *et al.,* 1985a; Suchey & Brooks, 1990, p.227; Ruengdit *et al.,* 2020, p.10). Biological processes generate various morphological aging pathways that exhibit distinct expressions of age in different anatomical regions. Therefore, a multifactorial method has been proposed as the most effective approach to account for these variations. (White & Folkens, 2005, p.363; Schanandore *et al.,* 2021, p.66, Navega *et al.,* 2022, p.18).

In the following sections, several widely used traditional methods are discussed. The accuracy of these traditional methods will later be compared with the performance of the multifactorial NN method. *Section 2.1.1* presents the Pubic symphysis method by Suchey and Brooks (1990), *section 2.1.2* introduces the method of the Auricular surface by Lovejoy *et al*., (1985b), *section 2.1.3* discusses the method for Cranial suture obliteration by Meindl and Lovejoy (1985), *section 2.1.4* will introduce the method of the Sternal rib ends by İşcan *et al.,* (1984). *Section 2.2* concludes the chapter with an assessment of the general performance and accuracy of these traditional methods.

2.1.1 Morphology of the Pubic Symphysis

Figure 1: Three stages of pubic symphysis morphology. Left: typical morphological pattern of young individuals. Middle: Pattern of intermediate individuals. Right: Typical pattern of elderly individuals. Figure from (Christensen et al., 2014, p.332, Figure 10.11)

The pubic symphysis is a cartilaginous joint that connects the two pubic bones of the os coxae of the pelvis. The pubic symphysis is crucial in load distribution of the pelvis as it is able to soften and stiffen to relieve stress from different parts of the pelvis (Ricci *et al.*, 2020, p.1,). This non-synovial joint consists of cartilage that synostoses with the passing of time (White & Folkens, 2005, p.374). The surface of the pubic symphysis contains grooves and billows at young adulthood which transform, with age, into a coarse indented surface (Fig.1).

Todd (1920) developed the first documentation of morphological changes when observing the Hamman-Todd collection of human remains. A ten-phase system was developed based on 306 (white) males with archival data (Todd, 1920, p.300; White & Folkens, 2005, p.375; Franklin, 2010, p.3). The observable morphological pattern of age difference between females and males was evident, but whether these changes were linked to chronological and biological age remained unknown (Brooks & Suchey, 1990, p.227,). Todd's method was extensively expanded upon because the sample Todd used mainly encompassed non-contemporary white male individuals resulting in a sample bias (Katz & Suchey, 1986, p.427; White & Folkens, 2005, p.375). With a modern sample the Suchey-Brooks method was created by including female samples and racial differences (Brooks & Suchey, 1990, p.227; White & Folkens, 2005, p.375).

As a result, a six-phase unisex version was developed, presenting general age morphologies that are similar in both sexes (Suchey & Brooks, 1990, p.232) (Fig.2). The first phase correlates with a

relatively young age group, featuring pubic symphysis morphology with relief and a developing rim. At the age of 35, the rim surrounding the surface completes, progressing further until the final phase, phase 6, where is broken down, and the relief degrades (White & Folkens, 2005, p.374). This method relies on the judgement of the osteologist or forensic anthropologist to categorize each pubic symphysis in one of the six phases corresponding to an age-category.

The Suchey-Brooks method has become one of the most widely used age-at-death methods (White & Folkens, 2005, p.375; Schanandore *et al.*, 2021, p.57). While employing multiple age estimation methods is beneficial for enhancing reliability, accuracy and reliability have varied across studies possibly due to differences in sample composition or interobserver error (Suchey & Brooks, 1990, p.227; Schanandore *et al.*, 2021, p.57). A meta-analysis of 18 studies including the Suchey-Brooks method found that it is paired with higher accuracy than other traditional methods, besides showing small accuracy differences between males and females, potentially because the method was developed with a sample featuring a higher percentage of males (Schanandore *et al.*, 202, p.63). Additionally, the pubic symphysis is frequently damaged in archaeological contexts, unfortunately limiting the applicability of this method.

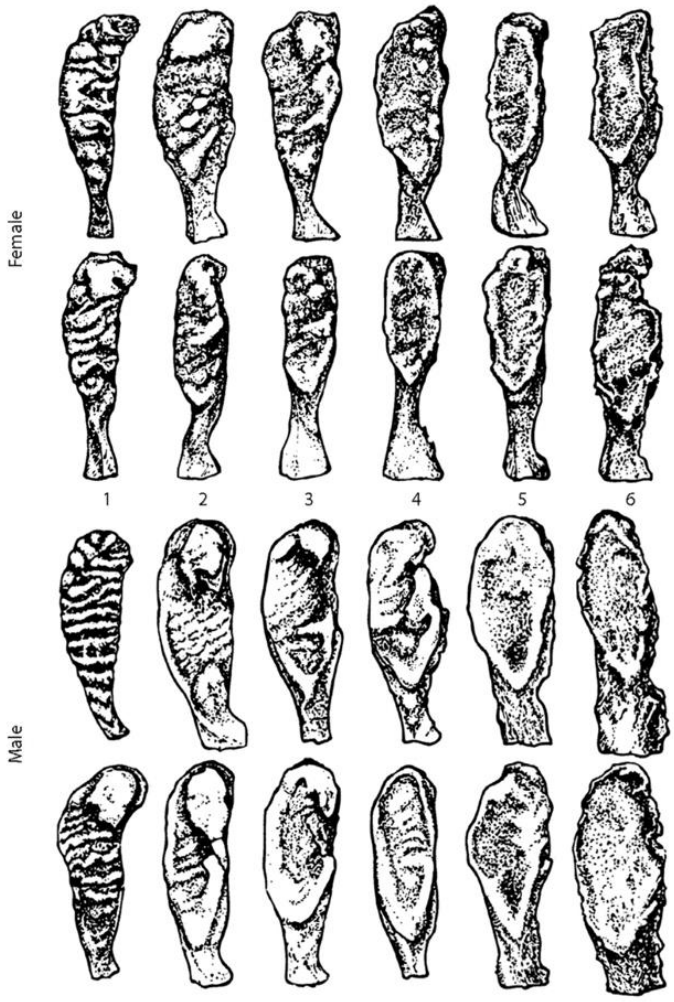


Figure 2: Morphological alterations in the pubic symphysis are assessed using the Suchey-Brooks system. Stages 1 to 6 express increasing age. Each phase is depicted in two expressions of the same age which are displayed in each column. Figure from (Buikstra & Ubelaker, 1994, p.23, Figure 7 & 8) drawings are made by Zbigniew Jastrzebski.

2.1.2 Auricular surface

Figure 3: Three stages of the morphology of the auricular surface. Left: the typical morphology pattern of young individuals. Middle: the pattern of intermediate individuals. Right: typical pattern of elderly individuals. Figure from (Christensen et al., 2014, p.334, Figure 10.13).

The auricular surface, shaped like an ear and located on the ilium, a part of the pelvis, contributes to the formation of the sacroiliac joint, connecting the pelvis to the sacrum. The texture created by cartilage and fibrous tissue is crucial, allowing slight movement while maintaining stability (Lovejoy *et al.*, 1985b, p.17). Compared to the pubic symphysis, the auricular surface has a higher chance of survival in the archaeological record (Lovejoy *et al.*, 1985b, p.1). Age-related changes in the auricular surface continue beyond the age of fifty, unlike the pubic symphysis, and both are claimed equally effective in estimating age accurately (Lovejoy *et al*., 1985b, p.15). The Suchey-Brooks method for age-at-death estimation is based on morphological changes in the pubic symphysis throughout life, divided into five phases corresponding to different age categories. These phases progress from an early post-epiphyseal phase characterized by a billowed surface, pronounced transverse organization, and fine granularity, to a young adult phase with less pronounced billowing and increased surface coarsening, and so forth, until the breakdown phase where the surface becomes porous and degrades (Fig. 3). While the Suchey-Brooks method can be a valuable tool for age-at-death estimation, it is important to note that accuracy and reliability can vary with the sample composition and the experience of the observer (Lovejoy *et al.*, 1985b, p.20).

While it is advantageous to categorize an auricular surface, there is a possibility that these surfaces may not precisely align with the descriptions, exhibiting features that could fit into more than one category. In such cases, the observer must determine which age phase corresponds to the overall age indication. Highlighting the importance of conducting multiple aging methods to enhance reliability. To minimize the risk of interobserver error, several osteologists were assigned to apply the system to the Todd collection. The resulting accurate age-at-death estimates demonstrate the replicability of this method (Lovejoy *et al.*, 1985b, p.27). Additionally, the data obtained from this test suggest that the auricular surface method is comparable in accuracy to the pubic symphysis method. This is particularly significant given that the auricular surface is more likely to be preserved in the archaeological record than the pubic symphysis, due to it being more exposed (Lovejoy *et al.*, 1985b, p.28).

2.1.3 Cranial suture closure

Figure 4: Stage progression of sutures. Top left: open suture. Top right: Minimally closed suture. Bottom left: partly obliterated suture. Bottom right: obliterated suture. Figure from (Christensen et al., 2014, p.340, Figure 10.18).

The cranial sutures consist of fibrous tissue that secures the cranial bones, and are acknowledged as potential indicators for age-at-death due to their progressive closure (Meindl & Lovejoy, 1985, p.57). Despite being utilized since the 16th century, the closure of sutures as a method for estimating age, faced criticism and was subsequently deemed unreliable by Brooks in 1955 (White & Folkens, 2005, p.369). During this period, the pursuit of discovering an accurate age estimation method led to the rejection of many methods that failed to deliver high accuracies (>80%) (Meindl & Lovejoy, 1985, p.57; Bailey & Vidoli, 2023, p.183). Meindl and Lovejoy (1985) had a preference for combining multiple methods to enhance overall precision which resulted in the revised method for suture closure in 1985 (p.57).

According to the methodology of Meindl and Lovejoy (1985), the ectocranial suture closure system is presented as a valuable age indicator, in which they claim similar performance to the Pubic symphysis (p.65). The cranial sutures are divided into segments, excluding sites showing no correlation to age, resulting in the lateral-anterior and vault suture systems comprising of 17 scorable segments (Fig.4). Each segment is assigned a score ranging from 0 (open) to 3 (obliterated) based on observed closure

levels (Fig.5). These scores are then combined, providing a composite score for both systems that represents an estimate of chronological age.

In summary, the main finding of this study is that the method is effective, particularly when complemented by other age estimation methods. The authors state that a method focusing on one particular anatomical region would not achieve 100% accuracy in estimating chronological age; rather, its strengths lie in the valuable information contributing to age-at-death estimation (Meindl & Lovejoy, 1985, p.65). This perspective remains relevant, as cranial suture closure proves relatively accurate when supported with other age indicators, with the expectation of further refinements through recent technological advancements (Ruengdit *et al.*, 2020, p.10).

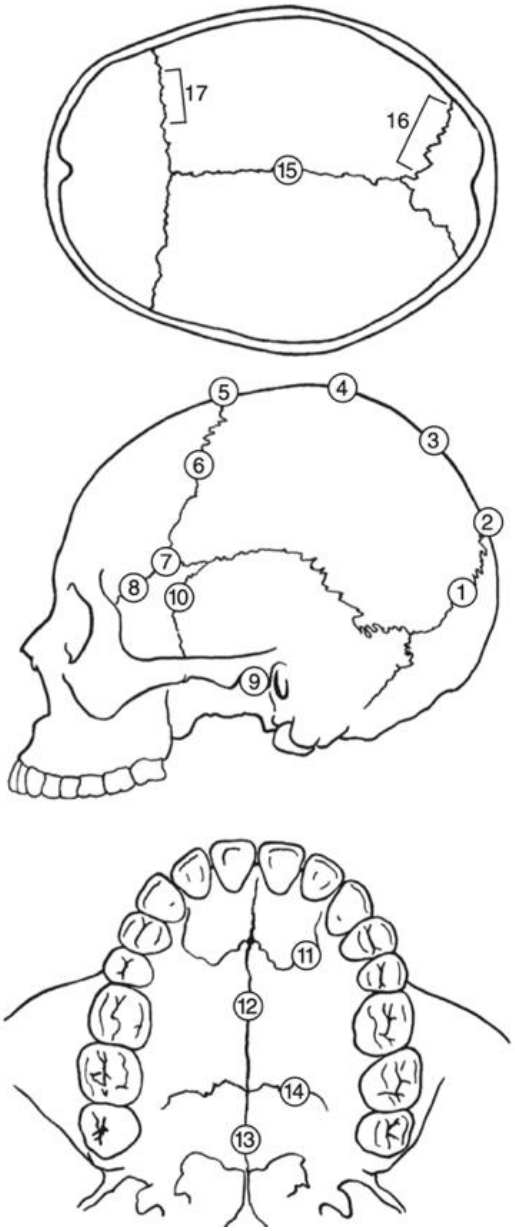


Figure 5: All sites where the suture obliteration score is given. Top: cranial view of the cranium. Middle: Sagittal view of the cranium. Bottom: caudal view of the cranium (maxillary and palatine bones). Figure from (White & Folkens, 2005, p.370, Figure 19.4).

2.1.4 Degradation of the Sternal ends of the fourth rib

Figure 6: Three degenerative stages of the sternal ends. Left: Typical appearance of young individuals. Middle: Appearance of intermediate individuals. Right: Typical appearance of elderly individuals. Figure from (Shook et al., 2019, p.562, Figure 15.15).

The coastal cartilage, or sternal rib ends, represent the distal portions of ribs articulating with the sternum, known as the true ribs, comprising the initial seven rib pairs. While the first and second ribs may display age-related changes more rapidly, the third to fifth ribs show no discernible morphological differences, reckoning them equally reliable for age-at-death estimation (İşcan *et al.*, 1984, p.155). Multiple studies support this view, emphasizing the nearly equivalent accuracy of ribs 3-5 in age estimation, with intercostal differences typically limited to one phase (Loth *et al.*, 1994, p.141; İşcan *et al.*, 1984, p.155). In addition, the first rib also introduces itself as a reliable indicator for age-at-death estimation and it can be used on its own (Kunos *et al.*, 1999, p.322; Luna & Aranda, 2022, p.2188). During adolescence, the sternal rib end displays a billowed appearance, transitioning into a hollow cup shape in middle-aged adults and further evolves into a deeper and more irregular form in old adults (Fig. 6). Radiographic investigations revealed age-related mineralization at the end of the fourth rib (İşcan *et al*., 1984, p.147).

Despite its potential, the sternal rib end received limited attention for age-at-death estimation until İşcan *et al.* (1984) proposed a method based on three components: pit depth, pit shape, and rim and wall configurations. Pit depth is measured from the cavity base to the highest point on the surrounding wall (İşcan *et al.*, 1984, p.148). The pit shape transforms, from a V-shape to a U-shape with age. The rim and wall configurations, evolving from a smooth to an irregular scalloped wall and eventually an irregular and sharp wall, are essential to the method (İşcan, 1984, p.152).

It is crucial to note that this method exclusively focused on male ribs and poses challenges in forensic and archaeological contexts were identifying or retrieving the fourth rib is often difficult (Franklin, 2010, p.4). While the effectiveness of this method is deemed comparable to the pubic symphysis, there is a significant risk of damage to the rib end over time and under various environmental conditions (İşcan et al., 1984, p.155).

2.2 Comparing traditional age-related features with degenerative traits in skeletal analysis

Age-related changes in the skeleton, occurring progressively with advancing age, are regarded as a natural part of the aging process. Most traditional age-at-death estimation methods are based on such constant changes which generally manifest in the same rate seen throughout all individuals in a skeletal collection and are not indicative of an underlying pathology. However, the appearance of these traits may vary due to genetic and environmental factors. Degenerative traits, often resulting from excessive pressure on the skeleton such as bone erosion, fractures, or pathological processes, can accompany aging, contributing to the natural degeneration of skeletons.

The drawback of employing this kind of parameter to estimate age-at-death is that there is a possibility that it does not represent the young individuals that do not yet display any degenerative traits (Navega *et al.,* 2022, p.20). In addition, the assessment of morphological features relies heavily on visual evaluation which is prone to subjectiveness of the investigator (Kotěrová *et al*., 2018, p.171). Degenerative alterations are more likely to vary among individuals and populations than developmental changes, resulting in an increased risk of bias (Boldsen *et al.,* 2002, p.75) Because of possible differences in degenerative traits between sex, pooled data has the ability to balance the misinterpretations and drawbacks from sex-specific methods out (Navega *et al.,* 2022, p.4). While some studies emphasize the advantage of utilizing sex-specific methods, there are studies that claim the opposite. Interestingly, Buk *et al.,* concluded that the knowledge of sex does not matter in age-at-death estimation (Buk *et al.,* 2012, p.8; Kotěrová *et al*., 2018, p.169). Transition analysis was mentioned before as a method to reduce age mimicry, but also aims to discern age-related variations in skeletal elements. The methodology that will be employed in this study strongly resembles a transition analysis approach that aims to capture the aging trajectories of skeletal features into corresponding stages. This multifactorial transition method will combine both degenerative and morphological features from multiple anatomical regions of the skeleton to retain an extensive result.

2.3 Conclusion regarding traditional age-at-death estimation methods

Age-at-death estimation is considered a crucial parameter in the identification of unknown skeletal remains, implying that broad age categories can significantly contribute to the identification process (Blau & Ubelaker, 2016, p.273). Traditional age-at-death estimation demands the use of multiple methods to yield the most precise outcome, recognizing that no single method can offer a precise and accurate chronological age estimate (Suchey & Brooks, 1990, p.237). The information obtained from each method holds value, many traditional methods depend on visual assessment, introducing the potential for interobserver error or subjective interpretation (Meindl & Lovejoy, 1985, p.65). Additionally, the majority of traditional methods were developed based on skeletal collections from specific time periods, potentially limiting their applicability to individuals from different contexts with comparable confidence (Ubelaker & Khosrowshahi, 2019, p.2). Practicing these methods on pre-historic skeletal remains could raise bias, as these methods were created with contemporary knowledge and materials. Age estimation in adults is further complicated due to the influence of individual factors, both environmental and genetic, on skeletal morphology, leading to bias. Furthermore, when developmental markers are absent, estimating age-at-death relies on observing

variable patterns of bone degradation (Franklin, 2010, p.3).

In the next section, the background off the NN will be established and further explained. This will hopefully gain deeper understanding of how a NN contributes to the study and what it is capable of.

2.4 Deep Randomized Neural networks

With the rapid rise of artificial intelligence, its applications are expanding simultaneously. For osteology and forensic anthropology, machine learning seems promising as skeletal data is often liable to subjective interpretation. The recent advances in machine learning and NNs have demonstrated promising perspectives in improving the accuracy and consistency of age-at-death estimation (Navega *et al.*, 2022, p.25). NNs are a subgroup of machine learning in the larger field of AI, which can learn a number of complex patterns and relationships from data, identifying what is difficult to detect using traditional age estimation methods and reducing the variability introduced by subjective interpretation. The utilization of NNs can provide understanding of these complex patterns of skeletal development which could lead to the creation of new age-at-death estimation methods.

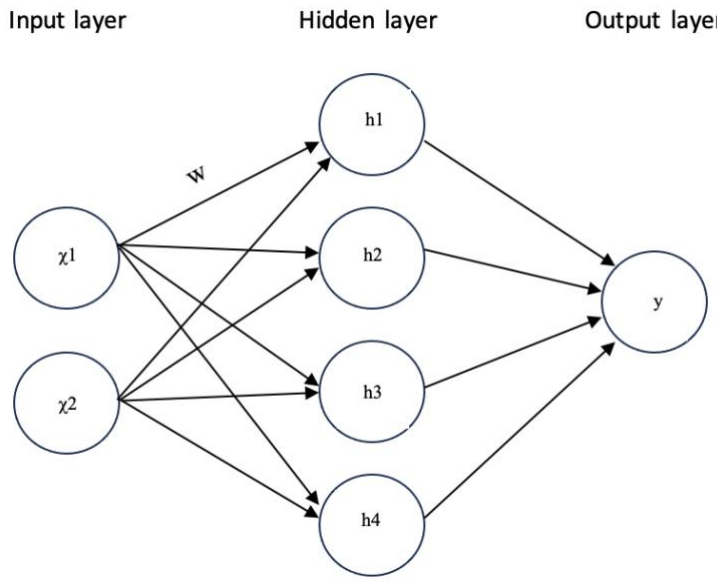


Figure 7: Schematic representation of the fundamental architecture of a neural network, wherein input layers denoted as (x) propagate signals through connections bearing respective weights (w) to the hidden layer. The hidden layer processes these signals and generates outputs based on the assigned weights, which are then conveyed to the output layer, represented as (y). Figure created by Babette Reus.

NNs are designed to mimic the interconnected networks of biological neurons in the human brain. They are structured in layers, forming a connected framework where these artificial neurons process data that is flowing from input to output (Aggarwal, 2018, p.17). Typically, most NNs are organized in three types of layers; an input and an output layer with intermediate layers in between, as seen in Figure 7 (Navega *et al.*, 2022, p.8). The depth of a NN refers to the number of intermediate layers present between input and output (Navega *et al.*, 2022, p.23). It is important to note that in Figure 7, only one intermediate layer is visible to provide a clear overview. However, often this is more complex and multiple intermediate layers exist, referred to as the “hidden” layers, which house

invisible (hidden) mathematical computations (Aggarwal, 2018, p.17). In the following sections, an attempt will be made to describe what occurs in these "hidden layers".

An artificial neuron is defined as a function that receives input and delivers output similar to the function:

$$y = f \cdot (x)$$

Where y = the age-at-death estimation and the output of the function and x = the skeletal trait, the function, "f" refers to the relationship between the input (x, skeletal traits) and the output (y, age-at-death estimation) (Navega *et al.*, 2022, p.8). This relationship could be a linear- or non-linear function, a regression model, decision tree, a NN etc.

Let us compare an artificial neuron to a biological neuron as displayed in Figure 8. In a NN, the inputs are similar to the electrical signals received by the dendrites of biological neurons, which are then transmitted to the synapses through the axon (Fig.8a). When a certain threshold potential is reached, the neuron fires (activates) and release neurotransmitters (chemical substances that act as messengers) from the synapses that act as the output and bind to the dendrites of the neighboring neuron, which can either inhibit it or activate it. In an artificial neuron (Fig.8b), these inputs (x) are assigned a weight (w) which represents the strength or importance of the connections similar to the strength of synaptic connections determining how much influence the signal has on the neighboring neuron (Gurney, 1997, p.13). When the weight of the input is higher, it indicates a stronger influence on the artificial neuron's output. Following the firing of an artificial neuron, its output is subsequently transmitted to other neurons within the network (Aggarwal, 2018, p.1). Comparable to the threshold potential determining neuron activation, the addition of an activation function (φ) establishes when a neuron activates and produces an output. Although there is a general disagreement on the comparison of the artificial and the biological neuron in machine learning, there is enough overlap to not disregard the similarities.

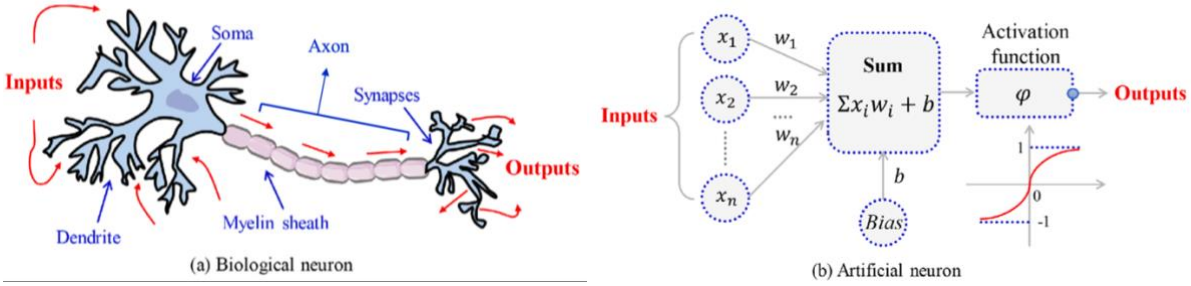


Figure 8: (a) Illustration of input through a biological neuron. (b) An illustration of how an artificial neuron processes an input to create an output. Figure adapted from (Wang et al., 2021, p.2284, Figure 5). Edited by Babette Reus

The weights play an important role in how a NN learns from data. When a NN is created, the weights in the input layer are assigned as random values. Each neuron in the following layer receives a weighted sum as its output from the preceding layer, which is a critical factor in determining activation based on whether it surpasses the threshold value (θ) (Fig.8b) (Aggarwal, 2018, p.29). This

threshold ranges from -1 to 1, and a value of >0 indicates activation, while a value of <0 indicates no activation. In a straightforward linear NN, surpassing the threshold is required for neuron activation, allowing the input to progress to the next layer. However, in more complex and non-linear NN, the application of an activation function (φ) becomes necessary (Fig.8b). The activation function is similar to an on/off switch and makes the discission whether a neuron should activate (Aggarwal, 2018, p.33). The threshold, ultimately decides if the neuron is activated or deactivated; activation occurs if the activation function surpasses the threshold value to determine if the neuron transmits its output to the next layer (Gurney, 1997, p.29). Now let us view this matter in a mathematical formula:

$$y = \sum_{i=1}^{n} x_i \, \omega_i$$

In the formula above, "*x*" symbolizes the input value, while "*w*" represents the value of the weight assigned to each neuron. The weights, equivalent to inhibitory and activating signals, may assume positive or negative values. The variable "*i*" stands for the index of input features, ranging from 1 to "n" (Gurney, 1997, p.29; Navega *et al.*, 2022, p.8). This formula imitates the mathematical computations occurring within the hidden layers.

$$y = \sum_{i=1}^{n} x_i \, \omega_i + b$$

In Figure 8b, a bias is mentioned which is not yet discussed. In the formula above, the bias (b) is added as a constant to the weighted sum of inputs before the activation function is applied. It is a fundamental part of the computation of a neuron because it allows the neuron to activate even when weighted sum is zero (Fig.8). The addition of bias can alter the effect of the activation function and forms a standard for activation, which influences how the neuron reacts to different inputs. Weights are the strengths of the connections and the bias is the indication of the neuron tends to be active or inactive (Aggarwal, 2018, p.2).

In the following section, the process of training a NN will be explained gradually and terms that are important to the study are introduced.

2.4.1 How a Neural network is trained

In the initial phases of NN training, weights and biases are assigned randomly and adjustments are made by making guesses. If the network computes an error, it adjusts the weights and biases just as it would adjust the strengths between the synapses in a biological neuron (Gurney, 1997, p.13; Aggarwal, 2018, p.2). Similar to the weights, the bias is also learned during the training process. As was mentioned in the previous section, the bias can alter the activation function. A positive bias shifts the threshold to the left, making the neuron more likely to activate, while a negative bias shifts the threshold to the right, decreasing activation (Gurney, 1997, p.64; Aggarwal, 2018, p.6). The network is first trained with known input and output pairs. For example, when the input data is an image, pixel values serve as input and the corresponding image description constitute the output. Training the

network involves predefined pairs of input images and correct output descriptions (Aggarwal, 2018, p.2). The NN aims to combine each input with the correct output and gives a predicted output.

The input flows then through the NN in a forward direction from input to output, layer by layer and results in an output, which is ultimately the prediction of the NN. In the following step, the predicted output is eventually compared to the known output and the difference is calculated with the loss-function. Subsequently, the predictions are then fed through the network again but this time, backwards, from the output to the input layer so that the NN can adjust the weights and biases accordingly to minimize the loss-function (Navega *et al.*, 2022, p.9). The phenomenon of information passing backwards through the network is referred to as backpropagation. The loss-function essentially is a measure of the network's performance, and it quantifies the relationship between the network's predictions and the desired outcome. The last step is the optimization phase, in which the weights and biases are updated using an optimization algorithm which purpose is to minimize loss. A commonly used optimization algorithm is gradient descent. Gradient descent involves calculating the lowest point along the gradient of the loss-function, which is equal to the lowest loss. This occurs in the direction of the steepest gradient, essentially terminating the process when the gradient is near zero. However, there is a possibility that the process stops at a local minimum, indicating that the loss is at a minimum but not necessarily the absolute minimum. Figure 9 illustrates this event, showcasing both the local minimum and the global minimum, which represents the true lowest point. To help overcome the algorithm not approaching the global minimum, running the algorithm from different points in the gradient of the loss-function could fasten the process.

When the loss is minimal, it is preferred to repeat the whole training process again for a few times, including all steps, to refine the NN's performance. In the following paragraph, general terms applying to NN training will be mentioned.

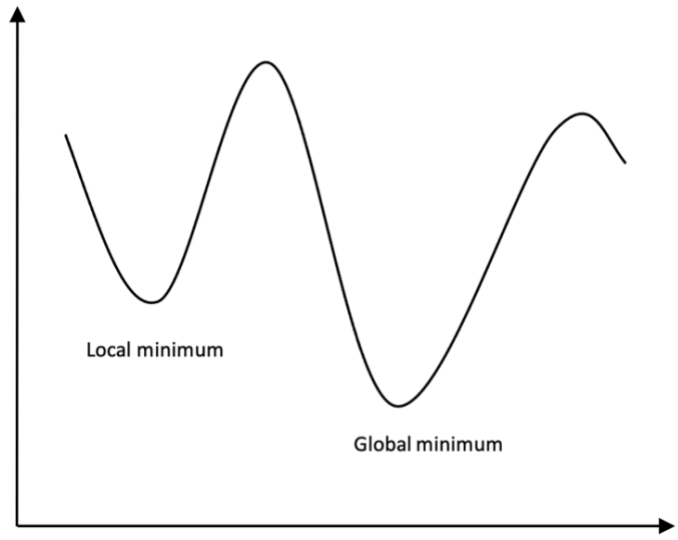


Figure 9: Visualization of the local- and global minimum in gradient descent. Figure by Babette Reus.

The term generalization refers to when a machine learning model learns from training data and becomes well adjusted at making predictions not only on the training data but also on new, similar

data it has never seen. Cross-validation on the other hand is a technique that assessess how well a model generelazes to unseen data. Cross-validation needs a dataset with both input data and corresponding known values and splits this dataset in a test (validation) set and a training set (Gurney, 1997, p.122). The model learns from the training set to make predictions on the test set (unseen data). What is special about Cross-validation is that it uses different portions of the datset to create new test and training sets, in this way, all data is used for training. It gives more insight in the performance of the model by testing it on different subsets of the dataset, ensuring that the algorithm is not solely memorizing the training data but actually learning useful patterns that can be applied to other problems.
Regularization is a term that defines training a model to prevent it from overfitting the training data and becoming too complex (Aggarwal, 2018, p.26). Overfitting occurs when a model imitates the training data perfectly and cannot generalize to unseen data anymore (Aggarwal, 2018, p.25). By adding a penalty to every individual weight in the network, it prevents the model from assigning large weights on any single feature (Aggarwal, 2018, p.181). Most often, every weight contributes based on the square of its magnitude defining the penalty as the sum of the squared weights across all layers of the network.

What could ultimately influence the age-at-death estimation is the quantity and quality of the skeletal data that is used to train and test the NN that will be utilized. The skeletal features encompassing the dataset, have to be relevant and preferably have a known association with age. In addition, data pre-processing is a valuable requirement to make the model learn patterns effectively. This includes normalization, cleaning and aiding to missing values etc. It is important that the training data is balanced regarding age-at-death categories because imbalanced data can lead to prediction bias. Cross-validation techniques and testing on separate datasets are essential to assess the generalization performance of the model. This helps in estimating how well the model will perform on unseen data. Regularization techniques avoid overfitting and further improve the generalization. In addition, it is essential to design the architecture of the NN, which encompass the number of hidden layers and the number of neurons in the network which can ultimately influence the ability to identify relationships in the data.

2.4.2 Current applications of AI in the osteological- and forensic anthropological field

The application of AI in the osteological and forensic anthropological field is at the beginning but there is an increasing awareness (Kotěrová *et al*., 2018, p.164). AI is not only utilized in age-at-death estimation but is claimed to be beneficial in sex estimation and even other implementations, which will be discussed in this section (Bewes *et al.*, 2019, p.42).
Buk *et al.* (2012) employed AI machine learning models to accurately age-at-death could be estimated from the pelvis (p.1). They assessed the surface of the pubic symphysis and sacro-pelvic region using a scoring method that describes the texture and morphology. They applied this method to 965 individuals from 9 different populations with known age. Unfortunately, they found that this method was inaccurate and could only assign individuals to robust age categories. They emphasized the necessity for additional measurable age indicators. Despite this, they expressed support for the role of artificial intelligence in advancing research in this field.
In contrast, Capella *et al.*, (2021), conducted a study where they employed machine learning to

identify commingled remains using osteometric data and 3D-scans (p.439). In this study they attempted to develop classification parameters to assign atlas and cranium to each corresponding individual. A total of 16 measurements were taken and used to train machine learning models with corresponding atlas and cranium. Unfortunately, this study deemed futile because the model could not assign each pair correctly. It is to note that experimental studies such as this one can lead to new innovations in the field. The thought of implementing AI in the problem of comingled remains seems insightful as this is seen as a problem with incomprehensible complexity. However, one cannot predict what new advancements in the future will bring to this issue.

AI is utilized for a spectrum of osteological problems, which applies to stature as well.

Czibula *et al.*, 2016, employed two machine learning based approaches for the problem of stature estimation in skeletal remains. As most of the traditional methods are based on bone measurements, this data was used as input for the machine learning algorithms (p.85). The performances of the two models were evaluated and compared with traditional methods in which the machine learning methods were able to outperform them.

Moreover, the estimation of age through neural networks is no novelty as this matter was already investigated in 2004, by Corsini *et al.* They tested the potential of NNs in estimating age-at-death on morphological changes in the auricular surface and pubic symphysis (p.163). The neural network displayed accurate results for the identification of younger individuals (20-29 years) and older individuals (>60 years). However, no accurate prediction was made for middle aged adults (30-59 years). This could possibly be the result of investigation only one anatomical region, the pelvis, or it could be that the ability of NN has improved since this time.

Navega *et al.* (2018), conducted a previous study in which they analyzed bone mineral density of the femur to estimate age-at-death on 100 Portuguese individuals with known age (p.497). This was investigated because bone mineral density decreases with age with a densitometer and a NN referred to as DXAGE that could generate age-at-death predictions from these variables. They found that the NN could reasonably estimate accurate with a MAE of ~9-13 years.

As was established in the introduction, Navega *et al.* (2022) conducted a study in which they created a multifactorial transition analysis which could score 101 morphological features on skeletal remains of 500 Portuguese individuals with known age. The scores were then implemented in a NN developed by the authors that is referred to as DRNNAGE in order to predict the age-at-death. The accuracy of the estimation by the NN was claimed to be over 95% with a MAE of ~6 years, making this a highly recommendable tool for implementation in other studies.

Furthermore, Rizos *et al.*, (2023) recently tested the accuracy of DRNNAGE software on 219 individuals of the Athens collection, a modern Greek skeletal sample. They unfortunately express some disappointment regarding DRNNAGE software but this will be further discussed in *Chapter 5*.

Štepanovský *et al*. (2023) proposed a data-mining model for adult age-at-death estimation based on 3D scans of the auricular surface based on 688 individuals from multiple populations (European and Asian) (p.1). To perform the 3D scanning, a high-resolution laser scanner was used and software was developed referred to as CoxAGE3D. What is interesting is that the method does not require any specific knowledge so the method is universally applicable. However, there is no confident result as the method can predict age-at-death with an MAE of 12.4 years (Pearson's r=0.56) (Štepanovský *et al*., 2023, p.7) but they claim to be comparable in performance with other traditional methods.

Toneva *et al.*, (2020) demonstrated that NNs can provide promising results when they are employed on cranial measurements to estimate sex. In this study, they tested three machine learning models,

including a NN which could accurately (>90%) estimate the sex of individuals (p.1). From the skeletal remains of 393 Bulgarian adults, 3D cranial models were created from which a set of 64 measurements and 22 indices on the cranium were fed in the NN. They showed that the NN could outperform traditional metric methods for sex estimation, further stimulating the potential AI models can offer in classification problems in osteology and forensic anthropology.

2.4.3 The potential of AI in the osteological- and forensic anthropological field

It is important to be aware of the potential bias in AI. When a NN is trained with predominantly similar data (i.e., white males), it may become less accurate when it is used to analyze the remains of unseen data (i.e., other racial or ethnic groups). To minimize the risk of bias, it is important to use diverse and representative datasets when training. Additionally, it is important to ensure that the AI model is transparent, so that datasets cannot be fabricated or affected by fraud. The use of AI in osteology and forensic anthropology could lead to job displacement if AI is used to automate tasks such as the analysis of skeletal remains. Despite the potential risks, AI also has the potential to provide valuable assistance in these fields when AI can be used to automate tasks such as data analysis and pre-processing and to identify patterns in data that would be difficult to detect. Because NNs are not susceptible to human biases, considered quite fast unlike the traditional methods (morphological and metric), and do not require extensive specialized knowledge to operate (Toneva *et al.*, 2020, p.14). As a result, identification of skeletal remains can be performed by a minimally trained osteologist in a relatively short time (Bewes *et al.*, 2019, p.42). AI is a rapidly developing field, and it is probable that AI will play an increasingly important part in osteology in the future. However, it is important to remember that AI is a tool and should be utilized in this manner to improve the work of osteologist and forensic anthropologists and not be a substitute for people. Furthermore, the interpretability of methods employing machine learning practices is a contemporary challenge that should not be overlooked (Navega *et al.*, 2022, p.24).

2.5 Conclusion of the chapter

The integration of AI in osteology and forensic anthropology has demonstrated significant promise, contributing to innovative solutions regarding identification challenges. The studies mentioned in *section 2.4* present the adaptability of AI applications, ranging from age-at-death prediction to issues like commingled remains. While certain studies exhibited comparable or superior accuracy to traditional methods (Czibula et al., 2016; Navega et al., 2018, 2022; Toneva et al., 2020), a degree of disappointment arises from the varying success observed in NN applications. Despite the nuanced outcomes, AI techniques are always recommended for further improvement. Continued research, method refinement, and interdisciplinary collaboration will expectedly contribute to further advancements in successful AI methods for skeletal identification in the future.
This chapter introduced background information regarding the purpose of this study. In the following chapter, the Methodology will be explained. *Section 3.1* will address more information regarding the MB11 collection and its distribution. *Section 3.2* addresses the strategy which will be presented through scoring tables. *Section 3.3* will cover how the training of the NN model was conducted, while *section 3.4* and *3.5* discuss the data pre-processing procedure and statistical analysis respectively.

Methodology

In this chapter, the methodology by Navega *et al.,* (2022) will be explained and how it was implemented in this study.

3.1 Data collection

In 2011, Hollandia, in co-operation with Leiden University conducted an excavation of the Keyser church in Middenbeemster, Netherlands (Hakvoort, 2013, p.9). Because of the construction for a new addition to the church, there was a good opportunity to document the surroundings and excavate the skeletal remains in the surrounding cemetery. The skeletal remains comprise individuals confined in coffins, originating from two distinct periods: one spanning from 1615 to 1829, and another from 1829 to 1866, with the majority of individuals belonging to the latter period (Hakvoort, 2013, p.35). The Laboratory of Human Osteoarcheology was the one responsible for conducting the identification research on this project. The assumption that the population was not of high social status can be made due to grave goods found (i.e., ceramics, glass, metal, bone). (Hakvoort, 2013, p.9). The MB11 collection consists of over 400 skeletons and is an archaeological collection. The individuals were not prosperous, indicated by the grave goods found alongside some of them, predominantly consisting of ceramics, metal, glass, and bone. Their overall health status was subpar, which while not unusual for the time period, suggested a general condition of poor health among these individuals (Hakvoort, 2013, p.9).

The skeletal collection that was employed in the study of Navega *et al.,* (2022), to train the NN model, is the Coimbra Identified Skeletal Collection (CISC) and the 21st Century Identified Skeletal Collection (XXI-ISC). Both collections are anatomical collections of which 500 human remains were sampled and investigated. All the individuals were of Portuguese origin whom died between 1904 and 2012 with an age range from 19-101. The reason information about this skeletal collection is implemented, is that it can be more useful for comparison with the MB11 collection as seen in Table 1.

Table 1: Demographic information about the data sampled from the MB11 collection and the CISC- and XXI-ISC collections studied in Navega et al., (2022).

Age-at-death	**Middenbeemster (BM11)**		**Pooled sex**	**CISC- XXI-ISC collections**		**Pooled sex**
	Female	**Male**		**Female**	**Male**	
Sample size	28	24	52	250	250	500
Mean (in years)	48.179	58.500	52.942	59.424	55.260	57.34
Min (in years)	21	19	19	19	19	19
Max (in years)	78	85	85	101	96	101
Std. deviation (in years)	17.885	20.720	19.733	23.556	22.141	22.93

The sample used for this study is derived from the portion of individuals that contained archival data from the MB11 (Table 1). The human remains that possessed documented archival data, were 118 in

total. However, due to time constraints, the research focused on a subset of these available skeletons, specifically 52 out of the total 118, in order to ensure an accurate and detailed analysis for each individual within the given timeframe. The subset of 52 skeletal remains exhibited an age distribution spanning from 19 to 85 years (52.942, ± 19.733 consisting of 28 females and 24 males (Fig.10). 101 skeletal features were scored using the macroscopic method introduced in Navega *et al.,* (2022, p.5). No individuals were excluded from the dataset due to taphonomy or pathological factors.

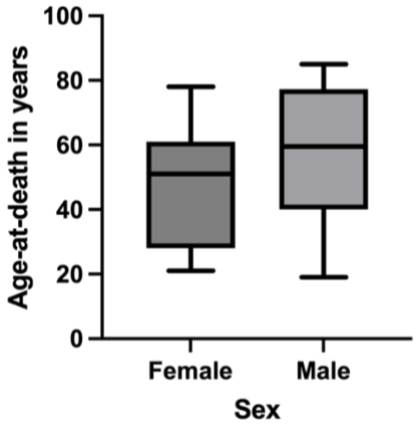


Figure 10: Age-at-death distribution of females and males in the MB11 collection of skeletal remains. Median with minimum and maximum age. Based off archival data.

3.2 Feature selection

Navega *et al*. (2022) proposed a macroscopic age-at-death estimation method for skeletal traits that are under investigated but can still provide valuable information as markers for age-at-death (Navega *et al.*, 2022, p.5). This comprises an easy applicable method that incorporates many features of different anatomical regions. The skeleton itself is seen as a marker instead of specific regions of the skeleton, which provides a more coherent connection between accurate age and estimated age-at-death. 64 unique skeletal traits were chosen that covered all the regions of the skeleton that show both developmental and degenerative aspects. The traits make up 101 when binary features are implemented (left and right, but the same feature. i.e., left humeral head and right humeral head). The features are all morphological do not exceed more than three stages of scoring. The stages can be either scored as 0, which implies no degeneration, 1 which implies degeneration or moderate degeneration and 2, which refers to substantial degeneration. When a feature cannot be scored, NA is noted. As the transition of each stage will vary between individuals, skeletal traits age in an essentially constant rate which implies that there is a general direction of senescent change (Boldsen *et al.*, 2002, p.74). Securing these changes in one transition stage to the next is rendered as an "transition analysis" estimation practice.

3.3 Strategy

The methodology as introduced in Navega *et al.,* (2022, p.5), will be slightly altered and presented in this section. The methodology is divided in the major anatomical regions: cranium, vertebrae, upper limbs, lower limbs, clavicle and first rib, pubic symphysis, sacroiliac joint, and the acetabulum.

3.3.1 Cranial scoring

The cranial scoring method is a modified version that is proposed in Boldsen *et al.* (2002) (Navega *et al*., 2022, p.5). The authors proposed an age-at-death method for the cranium and the pelvis but alternative morphological traits that display senescent changes can be employed (Boldsen *et al*., 2002, p.74). An advantage of this method is that it is already created in means of a five-stage, scoring-based system of age-related morphological variation and build upon traditional methods, such as the methods proposed by Todd (1920), McKern and Stewart (1957) and Meindl and Lovejoy (1985). Navega *et al*. (2022) modified this method solely for the cranial- and palatine sutures into a more compact two-stage system (p.5). In this methodology, all the sutural segments are scored according to the same suture stage description which can result in a given score of 0 (developmental) and 1 (degenerative) respectively (Table 2).

Table 2: Scoring description of cranial and palatine sutures. Each sutural segment is listed on the left and the stage description that applies to all segments is listed on the right.

Skeletal element	**Stage description**
Palatine sutures (posterior/median, transverse)	*Score 0*
Coronal-Sagittal suture (pars bregmatica)	There is a visible opening of the suture even though the bones could be juxtaposed and the suture narrow. However, there must be no signs of obliteration.
Coronal suture (pars pterica)	*Score 1*
Sagittal-Lambdoid suture (pars lambdica)	The suture shows signs of obliteration, which includes incomplete fusion or closure with the presence of bony junctions or it is completely obliterated.
Lambdoidal suture (pars asterica)	

3.3.2 Vertebrae scoring

To include morphological and degenerative traits, a three-stage scoring system was developed by Navega *et al.,* (2022), build upon previous research of Snodgrass *et al.,* (2004), Watanabe and Terazawa (2006) and Albert *et al.,* (2010). The methodology of Albert *et al.,* focuses on development and the fusion of the epiphyseal "ring" on the vertebral bodies and can estimate age-at-death accurately in an age range of approximately 14 to 24+ (Albert *et al.,* 2010, p.294). Furthermore, the fusion of the first and the second sacral vertebrae is one of the late-fusing segments and could stay

unfused for approximately thirty years of age which makes it a useful developmental indicator (Albert *et al.*, 2010, p.294). The methods of Snodgrass and Watanabe and Terazawa focus on degeneration of the vertebral bodies and osteophyte formation. It is implied that because of less variation in the lumbar vertebrae than in the thoracic vertebrae, the lumbar vertebrae contribute more to age estimation (Snodgrass *et al.*, 2004, p.3). In addition, the age-range that osteophyte formation covers is considerably, as osteophytes appear around thirty years of age but seems to be a more useful tool for elderly individuals (Watanabe & Terazawa, 2006, p.159). In this methodology, the inferior and superior surfaces of the third to the seventh cervical vertebrae and all the lumbar vertebrae are scored according to three-stage scoring system which can either be 0,1 or 2. In addition, the superior surface of the first sacral vertebrae is scored according to the stages of the cervical- and lumbar vertebrae. The fusion of the first- and second sacral vertebrae is scored according to a two-stage system which can either be 0 or 1. The description for the scoring of the cervical-, lumbar- and the first sacral vertebrae are displayed in Table 3. The descriptions for the fusion of the first- and second sacral vertebrae are displayed in Table 4.

Table 3: Scoring descriptions for the cervical, lumbar, and sacral vertebrae. Each vertebral element is listed on the left with a general stage description applying to all the vertebral elements on the right.

<table>
<tr><th>Skeletal element</th><th>Stage description</th></tr>
<tr><td>Cervical vertebra (C3, inferior surface)</td><td>Score 0</td></tr>
<tr><td>Cervical vertebra (C4, inferior- and superior surface)</td><td rowspan="2">There is no sign of degenerative change, the epiphyseal “ring” on the vertebral body is (partially) incomplete or fused and elevated. The surface of the body can display billows or grooves and is dense and compact.</td></tr>
<tr><td>Cervical vertebra (C5, inferior- and superior surface)</td></tr>
<tr><td>Cervical vertebra (C6, inferior- and superior surface)</td><td>Score 1</td></tr>
<tr><td>Cervical vertebra (C7, superior surface)</td><td rowspan="3">There are signs of degeneration. The margin of the vertebral body can be sharp and the surface of the body seem flattened and the “ring” could appear compressed. Microporosities can be present but in limited spatial distribution.</td></tr>
<tr><td>Lumbar vertebra (L1, inferior surface)</td></tr>
<tr><td>Lumbar vertebra (L2, inferior- and superior surface)</td></tr>
<tr><td>Lumbar vertebra (L3, inferior- and superior surface)</td><td>Score 2</td></tr>
<tr><td>Lumbar vertebra (L4, inferior- and superior surface)</td><td rowspan="3">The vertebra is degenerating. The margin of the vertebral body is sharp and lipped with a bony projection of at least 4 millimeters. The body of the vertebrae can be fused together by ossification of the vertebral ligaments. The surface of the body can be porous and irregular.</td></tr>
<tr><td>Lumbar vertebra (L5, superior surface)</td></tr>
<tr><td>First sacral segment (S1, superior surface)</td></tr>
</table>

Table 4: Scoring description for sacral fusion. The sacral element is listed on the left with the stage description on the right.

Skeletal element	Stage description
First and second sacral segment fusion (S1-S2)	*Score 0*
	The fusion of the first and second sacral body (S1-S2) is incomplete. The discontinuity on the anterior sacral surface is at least 1 centimeter.
	Score 1
	The fusion of the first and the second sacral body (S1-S2) is complete.

3.3.3 Upper limb and lower limb scoring

Features of physical activity and biomechanical stress are suggested to increase the accuracy of age-at-death. The general consensus was People thought these degenerative markers would only give a broad indication of age-at-death that can only determine old individuals from young individuals (Milner & Boldsen, 2012, p.227). However, age-at-death is an important factor in the appearance of degenerative traits (Navega *et al.*, 2022, p.6). It is difficult to base a scoring system on these traits as there are multiple skeletal elements that display different expressions. The methodology for scoring the upper- and lower limbs is built upon the traits in Buikstra & Ubelaker (1994) And Henderson *et al.*, (2012) whom developed standards for degenerative and entheseal changes. To revise a relatively simple system, the sole criteria are considered as absence or presence of degenerative traits which can be applied to joint- and musculoskeletal changes (Table 5). In addition, to increase accuracy in scoring for specific skeletal elements, a list of descriptions that make identifying stage 1 for these traits easier is listed in Table 6. If these specific elements do not align with these descriptions, then a score of 0 can be given.

Table 5: Scoring descriptions for the upper and lower limbs defined by general joint degradation and musculoskeletal degeneration. The skeletal element is listed on the left with the general stage description that applies to all the skeletal elements on the right.

Skeletal element	Stage description
Scapula glenoid fossa*	*Joint generation*
Humerus head*	*Score 0*
Humerus lesser tubercle	There are no signs of degenerative changes. The subchondral surface is dense and smooth while the joint margin is smooth.
Humerus greater tubercle	
Humerus capitulum and trochlea*	*Score 1*
Humerus medial epicondyle	There are signs of degenerative changes. Osteophytes on the joint margin are present that render the margin irregular. Porosities can be present on the margin as well. In addition, osteophytes and porosities may be present on the subchondral surface. The most severe case is the loss of articular morphology and eburnation.
Humerus lateral epicondyle	
Ulna proximal articular facet*	
Ulna olecranon	

Radius head*	*Musculoskeletal degeneration*
Radius tuberosity	*Score 0*
Os coxa iliac tuberosity	There are no signs of degenerative changes. The margin is smooth.
Os coxa ischial tuberosity	
Os coxa acetabulum*	*Score 1*
Femur head*	There are signs of degenerative changes. There can be two conditions that can both apply: 1) Osteophytes are present on the muscle attachment site contributing to an irregular appearance. 2) The bone surface displays irregularities, like granular texture, bony exostoses, erosions or cavitation.
Femur trochanteric fossa	
Femur greater trochanter	
Femur lesser trochanter	
Femur condyles*	
Tibia condyles*	
Patella articular surface*	
Patella base	
Calcaneus tuberosity	

*Table 6: Specific stage 1 descriptions that apply to features with an asterisk *.*

Skeletal element	**Stage 1 description**
Scapula glenoid fossa	The lipping on the margin of the glenoid fossa is pronounced and at least a third of the margin is affected.
Humerus head	The lipping on the margin of the humeral head is pronounced but less than in the glenoid fossa. It may take the form of a sharp elevated ring around the head or can be as severe as to take the shape of a collar.
Humerus capitulum and trochlea	Osteophytes on the margins of the articular surface might be present which is often accompanied by eburnation on the capitulum.
Ulna proximal articular facet	The lipping on the facets is pronounced but this is generally less than in other joints. Eburnation is uncommon on the articular facet.
Radius head	There is marginal lipping of the radial head and porosities on both the margin and the surface. Loss of bone density can be present in some cases.
Os coxa acetabulum	Osteophytes on the posterior column or inner margin of the acetabulum could be present which can damage the acetabular fossa. In addition, the acetabular fossa could display porosities, osteophytes and granular texture. In severe cases, there could be eburnation on the lunate surface of the acetabulum.
Femur head	The margin of the femoral head could display osteophytes. Bony nodules and an irregular surface around the fovea capitis could be present. In severe cases, the surface can be thin and an “osteophytic ring” can form around the femoral head giving it the appearance of a mushroom.
Femur condyles	There are signs of porosity on the articular surface of the condyles and marginal lipping. In more severe cases, eburnation is present on the articular surface of the condyles.
Tibia condyles	

Patella articular surface	There are signs of porosity on the articular surface of the patella and marginal lipping. In more severe cases, eburnation is present on the articular surface of the patella.

3.3.4 Clavicle and first rib scoring

Sternal epiphyseal fusion is aiding in the age estimation of younger individuals, as the epiphyses close around 30 years of age (Navega *et al.,* 2022, p,6). Falys and Prangle (2015) describe the morphological changes that occur post epiphyseal fusion and introduce a novel method to score these changes on osteophyte formation, porosity and surface topography (Falys & Prangle, 2015, p.203). As identifying elderly individuals from 40+ year individuals, this method can be considered of significance as porosity and surface topography can be suitable indicators while osteophyte formation plays less of a role (Falys & Prangle, 2015, p.213). For this methodology, a method is developed that encapsulates both developmental and degenerative changes in the clavicle. The morphological changes of the sternal end of certain ribs became an interesting method, but the drawbacks such as misidentifying the rib placement and dealing with damaged sternal ends in the archaeological record. Kunos *et al.* (1999) proposed a method that focused the first ribs' costal face, head and tubercle as the first rib has a distinctive morphology which can be easily identified and is often preserved (DiGangi et al., 2009, p.166). DiGangi *et al.* (2009) revised the method of Kunos *et al*. and created a scoring-based system that focused on the morphological traits that Kunos *et al.* used; the costal face and tubercle (DiGangi *et al.*, 2009, p.166). In this methodology, the method of Kunos *et al.* and DiGangi *et al.* is further improved upon to provide each skeletal element with their respective scoring system. For the sternal end of the clavicle and the costal face of the rib, a three-stage system was created which can be given scores 0,1 and 2. For the acromial end of the clavicle and the tubercle of the first rib this is a two-stage system which can be given scores 0 and 1 (Table 7).

Table 7: Scoring descriptions of the clavicle and first rib. The skeletal elements are listed on the left and the corresponding stage description on the right.

Skeletal element	**Stage description**
Clavicle sternal end	*Score 0*
	The epiphysis of the sternal end of the clavicle is unfused or partly fused.
	Score 1
	The epiphysis of the sternal end of the clavicle is fused. The surface of the sternal end has a smooth to granular texture. The margin displays no signs of osteophytes or irregularities. If porosities are present their spatial distribution is limited to less than one third of the surface.
	Score 2
	The sternal end of the clavicle is coarsely granular and marginal osteophytes might be present. Porosities occur and are present in more than half of the sternal surface.
Clavicle acromial end	*Score 0*
	The surface of the acromial end is smooth or is finely granular.
	Score 1

	Macro porosities are present on the acromial end of the clavicle. The surface may appear thin and trabecular bone can be visible.
1st rib costal face	*Score 0*
	The surface is flat and narrow, defined by a smooth texture and transverse ridges.
	Score 1
	The texture of the costal face is defined by an increasing cribriform pattern. The margin of the sternal end might become projected and scalloped with an increasing concavity.
	Score 2
	The margin of the sternal end has become a hollow shaft with an increasing concavity due to excessive ossification of cartilage. The surface is rugged and in severe cases sternocostal fusion can occur.
1st rib tubercle	*Score 0*
	The tubercle and its periarticular region are smooth.
	Score 1
	The articular surface of the tubercle is coarsely granular, porosities may be present and the margin can be lipped. The periarticular region is corrugated.

3.3.5 Pubic symphysis scoring

The pubic symphysis is the most used skeletal trait in age-at-death estimation (Navega *et al.*, 2022, p.7). As there are many methods existing that evolve around the pubic symphysis, such as scoring systems and casts made for comparison there is already sufficient research conducted on this matter. In this methodology, based upon the research of Suchey and Brooks (1985) and Todd (1920) a revised three-system scoring method is developed. This method tries to incorporate the developmental aspects and degenerative aspects of the symphyseal rim, topography and texture (Table 8).

Table 8: Scoring descriptions of the pubic symphysis. The skeletal elements are listed on the left and the corresponding stage description on the right.

Skeletal element	**Stage description**
Symphyseal rim	*Score 0*
	The rim is forming and is still incomplete. In early stages of rim formation, the rim is defined by the continuum between the face and the neighboring structures (pubic tubercle and pubic ramus).
	Score 1
	The formation of the rim is complete and is forms an elevated margin around the symphyseal phase. (Ventral hiatus, where the rim does not fully develop should not be confused with score 0 and 2).
	Score 2

	The rim is breaking down which can be accompanied with lipping and erosion, porosities and pitting of the dorsal and ventral margins.
Symphyseal topography	*Score 0*
	The topography is defined by a billowed surface of the symphyseal face.
	Score 1
	The topography is evolved from a billowed surface and becomes flattened and homogenous.
	Score 2
	The topography of the symphyseal face depresses and becomes irregular.
Symphyseal texture	*Score 0*
	The symphyseal texture is dense and smooth or finely grained.
	Score 1
	The symphyseal texture has become coarsely grained and micro porosities may occur in a limited spatial distribution.
	Score 2
	The symphyseal texture is less dense and eroded. Porosities occur along with bony formations

3.3.6 Sacroiliac joint scoring

Lovejoy *et al.* (1985b) and Buckberry and Chamberlain (2002) contributed much to the development of the methods for the auricular surface. For this methodology, their work was used to create a simple scoring system based on textural- and marginal changes. A method proposed by Passalacqua (2009) that describes a six-stage system for seven morphological changes on the auricular surface of the sacrum. This method both encompasses developmental- and degenerative aspects which can either be present or absent (Passalacqua, 2009, p.261). As the sacrum is found to be accessible for age-at-death estimation, it was utilized to create the methodology for this study and was revised into a two-system method. Described here, are a stage description of the general surface texture and marginal changes of the auricular surfaces of the ilium and the sacrum (Table 9).

Table 9: Scoring descriptions of the sacroiliac joint. The skeletal elements are listed on the left and the corresponding stage description on the right.

Skeletal element	**Stage description**
Iliac auricular surface texture	*Score 0*
	The surface is dense and smooth to finely granular with no porosities. Residual shallow billows might be visible.
	Score 1
	The surface transitions to a coarsely granular texture. Bony exostoses and micro porosities might occur, but in a limited spatial distribution.

	Score 2
	The surface becomes irregularly granular and eroded. Macro porosities are present in a clustered distribution.
Iliac auricular surface margin	*Score 0*
	The margin is pronounced and smooth.
	Score 1
	The margin is irregular, sharp or lipped.
Sacral auricular surface texture	*Score 0*
	The surface is dense and smooth to finely granular with no porosities. Residual shallow billows might be visible.
	Score 1
	The surface is coarsely granular. Porosities are present in a clustered distribution.
Sacral auricular surface margin	*Score 0*
	The margin is pronounced and smooth.
	Score 1
	The margin is irregular, sharp or lipped.

3.3.7 Acetabulum scoring

As there is no abundance in traditional methods that incorporate the acetabular surface, although it is a feature that is often intact in the archaeological record (San Millán *et al.,* 2016, p.23; Navega *et al.,* 2022, p.7). The method of Calce (2011), who revised the seven-stage method of Rissech *et al.* (2006) and decreased it to a three-stage scoring based method that was based on three features that had a high correlation with age: acetabular groove, apex activity and rim porosity (Calce, 2011, p.2). San Millán *et al.,* (2016) also revised the method of Rissech *et al.* (2006) and focusses on improving the scoring ability while making the method more applicable to both sexes (San Millán *et al.,* 2016, p.2). In this methodology, a three-system of three skeletal elements is developed based on the works of Calce and San Millán *et al.* (Table 10).

Table 10: Scoring descriptions of the acetabulum. The skeletal elements are listed on the left and the corresponding stage description on the right

Skeletal element	**Stage description**
Acetabular rim	*Score 0*
	The rim is pronounced and smooth. The acetabular wall presents no significant porosity.
	Score 1
	Osteophytes can be present on the rim defined by an osteophytic crest of approximately 1 millimeter. The rim is sharp and macro porosity

	and a rough surface might be present on the posterior wall and inferior iliac spine.
	Score 2
	The rim is eroded and irregular. An elevated osteophytic crest, more than four millimeters is present on the rim. Porosities and new bone formation can form on the lunate surface
Acetabular posterior horn	*Score 0*
	The apex is smooth with no osteophytes.
	Score 1
	The apex is sharp with an osteophyte of at least 2 millimeters.
	Score 2
	The apex contains an osteophyte of at least three millimeters. In severe cases, bone proliferation can create a bony bridge in the acetabular notch.
Acetabular fossa	*Score 0*
	The acetabular fossa is smooth and dense and there is no osteophytic activity.
	Score 1
	There are signs of degeneration. The edge of the fossa is rough and an osteophyte of at least one to three millimeters is present. The central surface can display porosities.
	Score 2
	The edge of the fossa can display osteophytic cresting which can partly obliterate the fossa in severe cases. The central surface has lost density and has become eroded at the point where trabecular bone might become visible.

3.4 Model training and experimental design

Navega and Cunha (2020) created a single layer NN to estimate age-at-death from data of the sacroiliac joint which was used as a base in this study. Gradient based learning is costly and needs technical knowledge to conduct thus the weights of the hidden layer where randomly assigned, through selecting random weights from a probability distribution (Navega *et al.,* 2022, p.9). Regularization is applied to the network and optimized by cross-validation (Navega *et al.,* 2022, p.10). However, applying this single layer network in age-at-death context, received critique that the network should be deeper (consist of more layers of neurons). To attend to the critique and improve the network, the current network was deepened according to Shi *et al.,* (2021) who proposed a method for deep randomized NN models. The network design is very similar to the single layer network but uses a randomized technique that allows a mathematical equation to be reused at along the depth of

the network instead of applying it just once for the final predictions. The first layer of the network receives input (skeletal features) and processes it through a mathematical function. This results in multiple intermediate age estimates, subsequent layers receive the result from the previous layer and apply the same process with their respective weights. By averaging all these a final age-at-death prediction is obtained. In this approach the network learns to collaborate across different depths, offering a unique way to improve prediction performance (Navega *et al.,* 2022, p.11). Each layer in the network is defined using mathematical equations.

To assess the performance of the network, cross-validation is employed. The dataset is repeatedly split into a training and a test (validation) set and trained multiple times (Navega *et al.,* 2022, p.12). Age-at-death was predicted with leave-one-out predictions on the test set which means that the entire model is trained except for one datapoint predicting the age for one data point while using the remaining data for training. The process is repeated for each missing datapoint ultill all the predictions are completed. In this study the process is repeated 1000 times and the dataset was split into 80% for training and 20% for testing in each repetition. The aim is to understand how well the models work under different conditions and levels of available data. Predictive intervals (95% PI) are computed to express the uncertainty in the model predictions which is performed by setting the uncertainty of a parameter (σ) to 0.05. (Navega *et al.,* 2022, p.13). The final network architecture consists of an eight-layer model with 32 neurons in each layer (Navega *et al.,* 2022, p.13).

3.5 Data preprocessing

Both sexes were pooled in this study because pooled models have the ability to balance out the drawbacks that sex-specific models are prone to, such as increased complexity and reduced sample size etc. (Navega *et al.,* 2022, p.4). Sex was not estimated during the data collection but retrieved from archival data as this method is generalized for both sexes. Unfortunately, there are numerous problems in multifactorial age-at-death estimation, which often result from the absence of data due to taphonomic factors or redundancy caused by bilateral data collection (Navega *et al.,* 2022, p.4). Despite the inherent asymmetry of the human body, one may hypothesize that the left and right sides exhibit negligible disparity (Navega *et al.,* 2022, p.4). To mitigate redundancy, the right score was chosen as a substitute when the left side was missing. Under this assumption, the left side was chosen as the source data, which brings the number of features to analyze from 101 to 64. This implies that only the left side of the skeleton was implemented in the data, unless data was missing which was then substituted with data from the right side of the skeleton.

For the remaining missing values, the nearest neighbor strategy was used. This strategy involved Jaccard similarity which measures the similarity between two datasets to substitute the missing values (Navega *et al.,* 2022, p.4). After the reducing the redundancy, Jaccard similarity was applied to the dataset using Phyton software[1]. Missing values made up 41.23% of the total entries but through pre-

[1] Python Software Foundation. Python Language Reference, version 3.12.0. Available at http://www.python.org

processing, the missing values could be brought down to 9.26% of the total entries. The data was then ready to be analyzed by DRNNAGE software[2] (α=0.05).

3.6 Statistical analysis

I followed the methodology outlined in *section 3.5*, and conducted the scoring of all (101) skeletal elements, with subsequent input of the scores into the DRNNAGE software, a deep randomized neural network created by the developers of Navega *et al.,* (2022). The correctness of the age-at-death estimate was determined when the estimated age provided by DRNNAGE fell within the 95% confidence intervals (CI) generated by the DRNNAGE software. The assessment of skeletal preservation state was explained through the percentage of skeletal elements that could be successfully scored. This percentage was calculated by dividing the number of scored skeletal elements by the total amount of skeletal features and multiplying the result by 100%. Additionally, a measure of bias was incorporated to denote the degree of over- or underestimation relative to archival age. This bias was quantified by subtracting the archival age from the estimated age. Statistical tests were performed with Phyton software. The mean absolute error was defined by the sum of absolute errors (bias) divided by the sample size. The correlation between archival age and estimated age was calculated with Spearman's rho.
To assess the effects of sex, age, and preservation, the data was divided into two datasets and tested for normality (Gaussian distribution). For the effects of three age categories, the dataset was divided into three groups of young age (25-40), intermediate age (41-60) and old age (>60) according to (Martrille *et al.*, 2007, p.302). The preservation score was expressed in the percentage of skeletal features that could be scored. Graphical figures were made using GraphPad Prism 9 software[3].

3.7 Conclusion of the chapter

In conclusion, the methodology employed in this study displays a comprehensive age-at-death estimation procedure. The incorporation of mostly less than three stages of scoring, ensures a straightforward system. The pre-processing of the raw data was conducted with Phyton software, and the aim was to mitigate redundancy and reduce missing values. After pre-processing the NN software DRNNAGE was used to convert the preprocessed data in age-at-death estimated. The statistical analysis was performed with Phyton as well and the aim was to test the data for significant patterns. Several statistical tests were performed including, the paired-t test, the Wilcoxon signed rank test, the one-way ANOVA, and Spearman's rho.

[2] Navega, David. (2022). DRNNAGE: Deep random NNs for adult skeletal age-at-death estimation. (0.0.1.0). Zenodo. https://doi.org/10.5281/zenodo.7433412

[3] GraphPad Prism version 9.5.0 for Mac, GraphPad Software, Boston, Massachusetts USA, www.graphpad.com".

Results

In this chapter, the outcomes of the investigation into the accuracy of age-at-death estimation with DRNNAGE software will be presented. Aligned with the research questions posed in *chapter 1*, a comprehensive analysis was conducted to explore the influence of various factors, including sex, age category, and preservation, on the DRNNAGE age-at-death estimate. *Section 4.1* highlighting differences between the estimated age-at-death and archival age, *section 4.2* examining gender-specific differences, *section 4.3* delving into age-category distinctions, and *chapter 4.4* exploring variations between individuals with low and higher preservation. *Section 4.5* concludes the chapter with a summary.

Table 11: All age-at-deaths estimated using NN (DRNNAGE software). Incorrect estimates are highlighted in red.

Individual ID	Sex	Percentage of the skeleton that could be scored (in %)	DRNNAGE-Age estimation (in years)	Archival data age (in years)	95% CI-lower bound (in years)	95% CI-upper bound (in years)	Individual bias
45/55	Female	84.158	**66.938**	47	51.969	83.484	19.938
47/45	Female	45.545	**25.615**	21	18.502	33.974	4.615
51/59	Male	73.257	**82.810**	74	64.297	100.449	8.810
53/290	Female	53.475	**63.974**	55	48.781	79.637	8.974
56/61	Female	34.653	**65.356**	78	48.827	82.460	-12.644
59/133	Male	93.069	**63.006**	38	48.011	78.424	25.006
60/37	Female	69.307	**25.692**	26	18.504	34.250	-0.308
77/98	Female	26.732	**69.473**	58	52.274	87.266	11.473
84/113	Female	85.149	**78.992**	52	61.475	97.189	26.992
88/94	Female	82.178	**78.560**	50	61.117	96.744	28.560
92/124	Male	64.356	**63.251**	59	48.120	78.795	4.251
93/126	Male	32.673	**68.941**	67	52.788	85.532	1.941
97/156	Female	84.158	**81.837**	78	63.502	99.644	3.837
100/159	Male	25.743	**83.669**	75	64.958	100.656	8.669
101/131	Female	83.168	**33.094**	39	23.349	43.315	-5.951
126/184	Female	14.851	**78.992**	67	61.475	97.189	11.992
137/491	Female	43.564	**59.066**	49	44.743	73.834	10.066
149/280	Female	79.208	**35.610**	25	25.392	46.273	10.610
151/666	Female	64.356	**33.387**	27	25.560	43.492	6.387
153/435	Male	63.366	**54.043**	57	40.566	67.889	-2.957
155/1509	Female	37.624	**55.006**	54	40.351	69.781	1.006
158/427	Male	87.129	**60.917**	60	46.393	75.935	0.917
160/613	Female	90.099	**26.485**	28	18.633	35.572	-1.515
162/316	Male	72.277	**58.911**	54	44.714	73.894	4.911
174/408	Female	50.495	**31.11**	45	21.987	40.567	-13.890
192/636	Female	12.871	**28.893**	53	20.282	37.939	-24.107

194/440	Male	65.347	**51.141**	58	38.274	64.535	-6.859
195/588	Female	42.572	**26.838**	54	18.963	35.462	-27.162
200/429	Male	4.950	**30.790**	85	21.551	40.608	-54.210
202/284	Female	22.772	**63.146**	28	47.034	79.343	35.146
213/220	Female	27.722	**54.834**	22	41.419	69.077	32.834
228/343	Male	10.891	**82.834**	78	64.277	100.458	4.834
236/335	Male	92.079	**24.413**	24	18.316	33.546	0.413
239/369	Male	76.238	**26.752**	23	18.908	35.3492	3.752
243/381	Female	69.307	**51.708**	62	38.746	65.202	-10.292
246/396	Male	64.356	**23.931**	19	18.273	33.367	4.931
250/402	Male	39.604	**64.472**	73	49.121	80.241	-8.528
253/466	Male	70.297	**78.807**	78	61.252	96.649	0.807
261/422	Male	26.733	**70.684**	79	54.542	87.692	-8.316
285/452	Male	78.218	**74.654**	71	57.826	92.313	3.684
289/477	Male	49.505	**49.989**	56	37.275	63.098	-6.011
294/487	Female	91.0891	**77.806**	66	60.208	95.719	11.806
297/498	Male	82.178	**78.558**	84	60.825	96.504	-5.442
302/509	Female	56.436	**71.293**	73	54.716	88.305	-1.707
303/520	Female	63.366	**43.225**	44	31.697	55.228	-0.775
306/561	Male	34.653	**38.309**	31	27.564	49.166	7.309
307/591	Female	65.347	**23.931**	21	18.273	33.367	2.931
309/616	Female	58.416	**33.735**	58	23.864	44.068	-24.265
310/550	Male	55.446	**33.342**	35	23.646	43.136	-1.658
313/926	Male	79.208	**37.517**	46	27.081	48.368	-8.483
317/649	Male	69.307	**70.594**	80	54.144	87.483	-9.406
319/669	Female	90.099	**69.186**	69	53.270	85.752	0.186

4.1 Estimated DRNNAGE age-at-death against archival age-at-death

The results of the age-at-death estimates by DRNNAGE are presented in Table 11. Out of all 52 age estimates 41 were correct (78.9%), which meant that the estimate fell in the 95% CI age range (Table 11) given by DRNNAGE software. Out of the 52 estimates, the DRNNAGE software tends to overestimate age 31 times (59.6%) and underestimates 21 times (40.4%). An example of the distribution of the estimated age and the corresponding 95% CI age range is displayed in Figure 11. The Wilcoxon signed rank test was performed between the estimated age and the archival age, because the assumption of a Gaussian distribution was not met ($p=0.283$, $\alpha=0.05$). The median difference between the groups was -1.474. The mean absolute error (MAE) was 10.423. Spearman's rho was calculated after the assumption of a Gaussian distribution was not met, which was significant ($r=0,7204$) as showed in Figure 12.

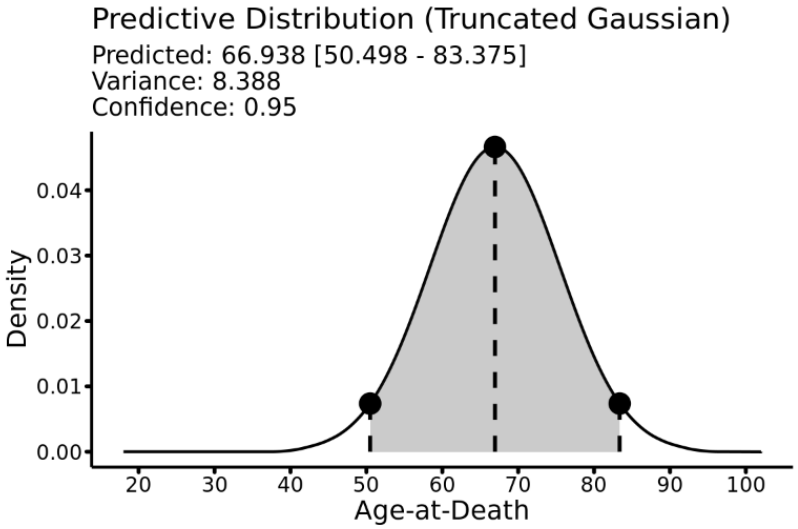


Figure 11: Probability distribution calculated using DRNNAGE. The probability distribution for individual 45/55 is displayed. The estimate is the dot in the middle with its corresponding 95% confidence interval age range showed in grey.

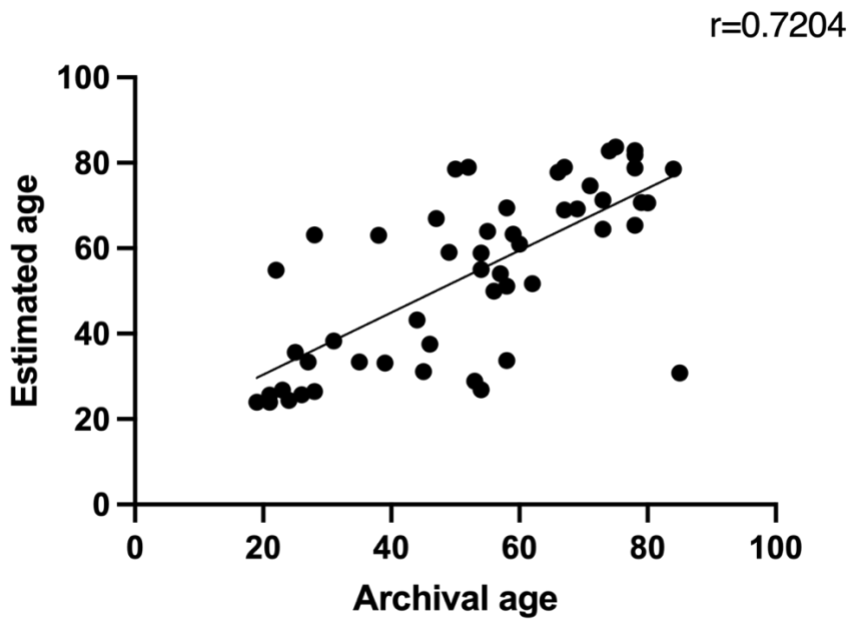


Figure 12: The DRNNAGE estimated age against the archival age with regression line r=0.7204.

4.2 Estimated DRNNAGE age-at-death difference between females and males

In the evaluation of all age-at-death estimates, it was observed that male individuals achieved a higher rate of accurate scoring compared to females. Male age-at-death estimates were correct for 22 out of the 24 individuals (91.7%) and female age-at-death estimates were correct for 19 out of the 28 individuals (67.9%). Females were overestimated in age 17 times (60.7%) while males where overestimated in age 14 times (58.3%). For females, the Wilcoxon signed rank test was performed

after the assumption of Gaussian distribution was not met (p=0.2270, α=0.05). The median difference between the groups was -3.384. For males, the paired-t test was performed (p=0.6400, α=0.05).

4.3 Estimated DRNNAGE age-at-death difference between age groups

Individuals that were fifty years or older were scored correctly 26 out of 32 times (81.3%) (Table 11). Individuals that were under fifty years old were scored correctly 15 out of 20 times (75%) (Table 11). For the group that was over fifty years or older, the paired-t test was performed after the assumptions had been met (p=0.5168, α=0.05). For the group that was under fifty years old, the paired-t test was performed after the assumptions had been met, and revealed a significant difference between the estimated age and the archival age (p=0.0343, α=0.05). The bias, the difference between the estimated age and the archival age, was included in Table 11, with 35.146 years being the highest overestimation and 54.210 years being the lowest underestimation. The highest and lowest bias are part of the group over fifty years in age (Fig.13). However, the bias of the group of fifty years and older is more concentrated along the zero-axis then the group under fifty which displays a more dispersed pattern.
To determine if the difference between young- (20-29 years), intermediate- (30-59 years), and old age groups (>60 years) have an effect on the age estimate, one-way analysis of variance (ANOVA) was performed (f=2.25, p=0.1160, α=0.05).

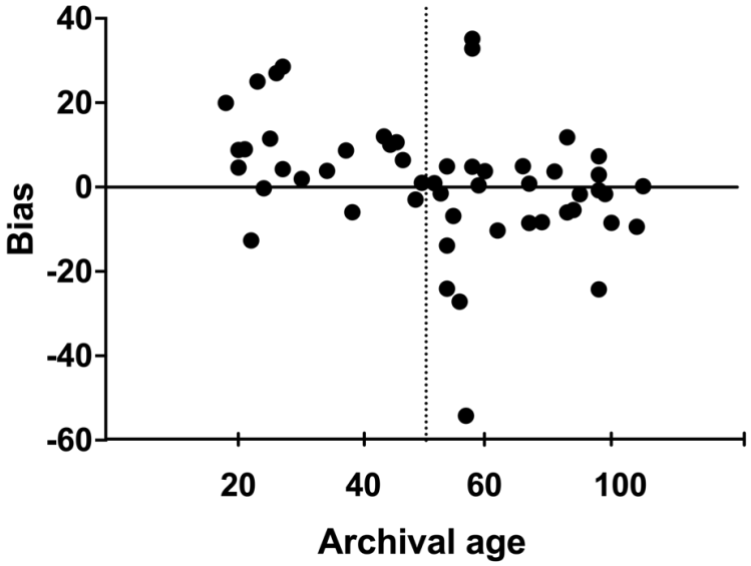


Figure 13: Archival age against the bias of the DRNNAGE age estimation, which can be either negative (underestimation) or positive (overestimation). The dotted line separates individuals over fifty from individuals under fifty.

4.4 Estimated DRNNAGE age-at-death against preservation score

The highest preservation score was 93.1%, while the lowest preservation score was 5%. 18 of the 52 individuals had a preservation score that was lower than fifty percent, while 34 individuals had a preservation score that was higher than fifty percent (Table 11). For the group with a preservation

score less than fifty percent, a paired-t test was performed ($p=0.9021$, $\alpha=0.05$). For the group with a preservation score higher than fifty percent, the Wilcoxon signed rank test was performed after the assumption of Gaussian distribution was not met ($p=0.2702$, $\alpha=0.05$).

4.5 Conclusion of the chapter

In summary, this investigation found that the DRNNAGE software can estimate age with approximately 78.9% accuracy. Males were more likely to be estimated correctly over females. In addition, individuals that were over fifty years, were scored correctly more often than individuals under fifty years of age. Finally, there was no significant difference found between preservation and the accuracy of the estimated age-at-death by DRNNAGE. These results will be further examined and discussed in the following discussion section, where the implications and potential explanations for the observed patterns will be explored.

Discussion

In this chapter, the transition from result presenting the results to a critical analysis of their significance will be important. This section contextualizes the findings within the study's hypotheses, exploring how the observed outcomes align with or challenge initial expectations. Emphasis is placed on linking the results with existing literature and addressing limitations transparently. By doing so, the intention is to provide a comprehensive understanding of the contributions and implications of this study within the osteological and forensic anthropological field. In the subsequent *section 5.1*, a summary of the main results will be provided, followed by the discussion and interpretations of these results in *section 5.2*. *Section 5.3* will address the study's research questions, while *section 5.4* will refer to the limitations of the study and the chapter will be concluded thereafter.

5.1 Summary of the results

The statistical analyses suggest a failure to reject the null hypothesis, indicating no notable distinction between the estimated age by DRNNAGE and archival age. Moreover, the DRNNAGE software displays predictive accuracy within the 95% CI in 41 out of 52 instances, implying an estimation accuracy of 78.9%. Spearman's rho reveals a significant correlation between the estimated age and the archival age ($r=0.7204$) and the mean absolute error is computed at 10.423. Specifically, in the case of females, there is insufficient evidence supporting a difference between these two variables. Similarly, for males, there is no statistical basis to infer a distinction between archival age and estimated age. Among those aged fifty years or older, no disparities emerge between estimated and archival ages. In contrast, the subgroup under fifty years manifests a statistically significant difference ($p<0.05$, $\alpha=0.05$). Regarding the preservation score, no significant difference is found between archival age and estimated age.

5.2 Interpretation of the results

When confronted with a p-value that is not significant, the data does not immediately ascertain that there is no difference between the observations; rather, the data fails to supply compelling evidence for such a distinction. It can only be suggested that the estimated age and the archival age are correlated. After performing the Spearman's rho to assess the strength of this correlation, a significant positive association was observed between the estimated age using the DRNNAGE software and the archival age ($r=0.7204$). Navega *et al.* (2022) claim a high accuracy of the age-at-death estimated by the DRNNAGE software, with a MAE of approximately 6 years across the adult lifespan (p.1). However, in this study, the performance was nuanced in the (MB11) Dutch medieval sample overall, with a MAE of 10.423 years and estimates deviating from the archival age as much as 54.210 years.

In cases where both p-values are not significant, as observed in both the female and male groups, the failure to reject the null hypothesis for both sexes suggest a correlation in age estimation accuracy between estimated and known ages. A contrasting finding emerges when comparing these results to the study conducted by Rizos *et al.* (2023), where females exhibited a higher correctness rate than males (52% vs. 41.3%) (p.4). However, this trend is not observed in the present study, indicating a

contrasting outcome where males are more frequently correctly estimated. This discrepancy might be attributed to sex-specific differences, such as females experiencing greater bone loss at a younger age than males (Agarwal & Grynpas, 2009, p.250). Furthermore, females exhibit greater variability in osteophyte formation in the vertebrae, a characteristic that might extend to other skeletal regions as well (Snodgrass, 2004, p.6). In the human postcranial skeleton, distinct variations exist in how sexual differences in growth rate and duration contribute to adult sexual dimorphism (Humphrey, 1998, p.72).

Of interest is the observation that the DRNNAGE software appears to show improved performance for individuals aged fifty or more than fifty years, as the estimated age-at-death significantly differed from the archival age in the group under fifty years. This implies that the NN model is unable to accurately predict age-at-death estimates that closely align with the archival age. These findings align with the study of Rizos *et al.* (2023) where DRNNAGE software was tested on a modern Greek sample, revealing a high accuracy on individuals over fifty years (p.8).
This variation could be attributed to the overrepresentation of individuals over fifty years in the MB11 collection sample, constituting 32 out of the 52 individuals (Fig.10). Similarly, the training datasets for the DRNNAGE software, namely the CISC and XX-ISC populations used by Navega et al. (2022), display a predominant distribution of individuals over fifty years of age (p.4). This pattern is also evident in the Athens Collection, as highlighted in the study by Rizos *et al.* (2023), where a substantial proportion of individuals surpass fifty years of age (p.2).

Concerning preservation, no significant difference was observed in the state of preservation, as indicated by the percentage of scorable features, and the age-at-death estimation. Although preservation can influence age estimation, given the vulnerability of degenerative features such as fragile sharp edges and osteophytes, it is unsurprising that the elderly are more susceptible to taphonomic changes (Ubelaker & Khosrowshahi, 2019, p.3). This suggests that the method exhibits significant performance for the elderly, even when dealing with skeletons impacted by taphonomic alterations commonly encountered in archaeological contexts.

5.3 Research questions

- This thesis aims to determine whether the trained deep random NN (DRNNAGE software) can accurately estimate the age-at-death of adult skeletal remains from a Dutch medieval sample.

The DRNNAGE software demonstrates considerate accuracy in predicting age-at-death, achieving correct estimations in 78.9% of cases. Its applicability appears robust in archaeological contexts, particularly within European populations, as aligned with the Portuguese and Greek populations that were tested by the NN model (Navega *et al.*, 2022, p.1; Rizos *et al.*, 2023, p.1). The significant positive correlation coefficient (r=0.7204) underscores the NN model's ability to align with archival age. However, the MAE of 10.423 years, in contrast to the approximately 6 years reported by Navega *et al*. (2023), suggests nuanced performance on the Dutch medieval skeletal sample. Notably, the model's accuracy appears promising for individuals over fifty years of age.

- Which factors can influence age-at-death estimation using deep random NNs?

Identifying the factors influencing age-at-death estimation through deep random neural networks, such as DRNNAGE, proves challenging. This study highlights age category (≥50 or <50 years) as a factor with substantial influence, overshadowing the impact of other potential factors. The absence of significant results for the tested factors prevents drawing conclusions. Consistent with findings from various studies (Rizos *et al.,* 2023, p.7; Lovejoy *et al.,* 1985a, p.12; Bedford *et al.,* 1993, p.287), the multifactorial approach employed by Navega *et al.* (2022) underscores the potential differential contribution of each anatomical region to the final age-at-death estimate. Notably, Rizos *et al.* (2023) evaluated the validity of each anatomical region outlined when replicating the methodology of Navega *et al.* (2022). Thereby concluding that cranial sutures exhibited the highest validity, followed by the clavicle and first rib, acetabulum, and pubic symphysis (p.7). Conversely, the vertebrae emerged as the region with the lowest validity.

- How does the accuracy of trained deep random NNs compare with the traditional methods of age-at-death estimation in adult skeletal remains?

The Suchey and Brooks (1990) method for age estimation, particularly focusing on the pubic symphysis, stands out as the preferred and most accurate traditional approach, followed by the auricular surface of the sacroiliac joint and the sternal ends of the fourth rib, as suggested by Bartelink (2019, p.330). Schanandore *et al.* (2021) conducted a meta-analysis of 18 studies on the Suchey and Brooks method, revealing a significant Spearman's correlation for both sexes combined (r=0.62) (p.56).
To compare the skeletal collection's performance, in a study by Sluis *et al.* (2022), three different methods for osteophyte formation were tested on 88 individuals of the MB11 collection, resulting in an accuracy range of 72.73% to 76.14% (p.1). Notably, the methodology for the vertebrae by Navega *et al.* (2022) draws significantly from two of the methods they employed, namely Watanabe and Terazawa (2006) and Snodgrass (2004) (Ch. 3.3.2). Comparing these findings, the age estimation accuracy of the DRNNAGE software appears slightly higher than reported by Sluis *et al.* (2022). However, this could be attributed to the multifactorial advantage of the DRNNAGE method, incorporating more valuable age estimation factors than the vertebrae alone. Traditional methods, often used in combination with others, may not fully capture the complexity of age estimation (Ubelaker & Khosrowshahi, 2019, p.2). Interestingly, the vertebrae demonstrated the lowest validity at 42.6% (Rizos *et al.*, 2023, p.7), suggesting that modifications by Navega *et al.* have made scoring for the vertebrae less reliable.
In contrast to other multifactorial methods, Bedford *et al.* (1993) tested regions such as the pubic symphysis, auricular surface, femur, and clavicle on a Canadian sample, yielding an MAE of 8.7 years (p.287). Lovejoy *et al.* (1985a) examined a multifactorial method involving the pubic symphysis, auricular surface, femur, dental wear, and suture closure, revealing a correlation above 0.72 (Pearson) with age (p.9).
While Navega *et al.* (2022) suggest their method should achieve an AME of approximately 6 years and an accuracy of 95%, the present study reports an AME of 10.423 years, accuracy of 78.9%, and a correlation of 0.7204 (Spearman's rho). This suggests that the performance of DRNNAGE is

comparable to other multifactorial methods in the field but falls short of the high accuracy rates often claimed by studies utilizing AI. When the method is compared to other AI-embedded methods to predict age-at-death, it shows similar performance (Corsini *et al*., 2004; Navega *et al.,* 2018; Štepanovský, *et al.,* 2023). However, there could be a discrepancy between the different aspects of skeletal identification as sex estimation proves to me more accurate (Toneva *et al.,* 2020, p.1).

- How can trained deep random NNs be used to improve age-at-death estimation in diverse populations and contexts?

To enhance the effectiveness of the DRNNAGE software, which is currently trained on data from the Portuguese CISC and XX-ISC populations, it is advisable to expand the training dataset to include more diverse populations. This can be achieved by incorporating datasets representing various ethnicities and regions, ensuring a broader representation of senescent features. If direct access to diverse datasets is challenging, an alternative approach involves training the model with datasets closely resembling the features required by traditional methods from diverse populations. Furthermore, conducting cross-validation studies across different populations can validate the method's accuracy in diverse contexts and improve its generalizability. To explore this aspect further, there is potential to adapt multifactorial methods regionally, tailoring them to specific population features and ethnic groups.
Variables related to senescent (aging) changes demonstrate a substantial impact on the success of classification models, surpassing the influence of other variables describing population characteristics. This implies that the aging process varies among populations, and these variations are reflected in model outcomes. Emphasizing the significance of input data, the origin of individuals remains crucial for accurate age estimation, particularly in studies where senescent changes play a significant role (Buk *et al.,* 2012, p.294). Recognizing the neglected differences in populations tested by studies on homogenous groups, it is essential for method development to consider the diverse nature of populations. Anticipating these variations in advance can lead to adjusted models, as different populations can show diverse rates of aging (Blau *et al.* p.282).
Considerations of population differences are crucial, as highlighted by distinctions between Portuguese and Dutch populations, as well as the contrast between anatomical and archaeological samples. Lovejoy *et al.,* (1985a, p.12) suggests that age determination in archaeological populations tends to be more accurate due to greater uniformity in environmental and genetic variables. This underscores the significance of diverse samples, encompassing various regions and ethnic backgrounds, as emphasized in the conclusion. The importance of diversity, as mentioned by Buk *et al.* (p.8), serves as a safeguard to ensure the applicability and reliability of age estimation methods across samples of unknown origin.

5.4 Limitations of the study

Sex-specific methods are prefered in age-at-death estimation as sexual dimorphism is often regarded as something which cannot be uniform (Humphrey, 1998, p.57). It is suggested that because of the different rates of aging between males and females, it is necessary to create independent scoring stages that can apply to multiple regions of the skeleton (Blau *et al.,* 2009, p.282). Each age

estimation method operates on a different scale, with some methods unaffected by sex, while others depend on the methodology and scoring of features (Blau *et al.*, 2009, p.282). Therefore, every study introducing or testing an age estimation method should assess the potential for sexual dimorphism (Kotěrová *et al.*, 2018, p.169).

In addition, this method is a multifactorial approach which is deemed superior over independent indicators (Martrille *et al.*, 2007, p.302). However, Rizos *et al.*, (2023), observed the lowest accuracy when all anatomical regions were combined to form an age estimate, rather than assessing them independently (p.8). This could be explained by the different aging trajectories influenced by internal and external factors within each anatomical region. Some skepticism toward multifactorial methods exists, suggesting that these approaches introduce unnecessary complexity without outperforming averaging multiple traditional methods or independent indicators. (Latham *et al.*, 2010, p.243) Since the vertebrae displayed the lowest validity, it might be helpful to incorporate the findings of Sluis *et al.*, (2022) to refine the methodology of Navega *et al.*, (2022) and hopefully make the accuracy of this anatomical region higher. Furthermore, it is suggested that the use of morphological and degenerative indicators limit estimate age-at-death estimation to the three broad categories: young adults, intermediate, and older individuals. To achieve more accurate age estimates, replacing visual scoring with objective methods, such as AI techniques, in extensive multi-population datasets is recommended (Buk *et al.*, 2012, p.8).

5.4.1 Problems regarding the Neural Network

An interesting pattern emerges in the data, revealing clusters of incorrect estimates. Table 11 suggests a potential clustering of estimated ages within the middle portion, spanning from individual 174/408 to individual 213/220. This clustering raises inquiries about potential influencing factors, notably the utilization of the nearest neighbor processing technique. Notably, the preservation scores within this clustered incorrect portion fall below 50.5%, indicating a limited opportunity for neighboring scores to influence each other during preprocessing.
In the context of NNs, a recurrent bias pattern is an inclination towards overestimating the ages of younger individuals and underestimating those of older individuals (Navega *et al.*, 2022, p.14). As the model is inclined to overestimate individuals instead of underestimate (Ch. 4.1), this is an interesting finding as this happens regardless of young or old individuals (Fig.10). There seems to be no difference in over- and underestimation of females and males. An explanation for this phenomenon could be that there is most of the times an underrepresentation of younger individuals in a skeletal population dataset, as seen in the sample of MB11, 20 out of 52 individuals are younger than fifty. This means that the NN is mostly trained on data of older individuals which makes it prone to overage on unseen data.

5.4.2 Effects of age group

It is notable that seven out of eleven inaccurate estimates fall within the intermediate age group (41-60 years), while three are within the young age group (25-40 years), and one in the old age group (>60 years) as indicated by (Martrille *et al.*, 2007, p.302). However, after performing a one-way ANOVA,

there appears to be no significant difference between the groups. The methodology of this study, which primarily focuses on developmental and degenerative features, introduces increased complexity in scoring intermediate stage transitions, potentially contributing to inaccuracies in estimation. The neural network exhibited commendable accuracy in identifying both younger individuals (20-29 years) and older individuals (>60 years). Conversely, accuracy was lacking in predicting ages for middle-aged adults (30-59 years). In addition, younger individuals are underrepresented in the MB collection sample (Fig.9). This inconsistency in representation raises questions about potential biases in age estimation models and the need for a more extensive representation of diverse age groups in the training datasets.

5.4.3 Effects of preservation, taphonomic, and diagenetic change

The assessment of degenerative changes introduces potential confusion with poor preservation, exemplified in Individual 246/396, where the presence of epiphyses on the humerus head and femur condyles initially suggested a very young individual. However, due to significant bone deterioration and loss, the final age estimation reached 24 years, surpassing certain other predictions made by the neural network. This prompts speculation that age overestimation may occur when dealing with incomplete skeletal elements, particularly as the model lacks specific training in subadult traits.

In the case of Individual 309/616, the presence of potential Pott's disease adds complexity to the interpretation, introducing the possibility of inaccuracies, especially when assessed by individuals with limited osteological experience. This complexity is reflected in the incorrect estimation of this individual, as indicated in Table 11. Consequently, heightened caution is recommended when dealing with individuals exhibiting various pathologies that can impact bone structures.

5.4.4 Effects of intra observer error

Concerning the impact of inter-observer error, Navega *et al.* (2022) contend that the methodology's interobserver error is negligible, citing a high concordance coefficient of 0.907 (Navega *et al.*, 2022, p.15). They attribute this result to their straightforward scoring strategy, which involves no more than three stages. However, in contrast to this view, the perspective gained from implementing the methodology in this study suggests that the descriptions for each skeletal feature were relatively brief and vague, leaving room for subjective interpretation. Rizos *et al.* (2023), who assessed the interobserver error following the scoring strategy Navega *et al.* (2022), reported a concordance coefficient of 0.717 and 0.748 for two observers (Rizos *et al.*, 2023, p.7). Such values are considered low for a method relying significantly on descriptions involving two or three stages.
Furthermore, it is important to note that the overall variability in skeletal morphology is only attributed to a limited extent by aging, therefore it is suggested that visual assessment of features introduces significant noise, which ultimately requires the need for an objective substitute for this process (Buk *et al.*, 2012, p.8).
In addition, as all the scoring work was done by me, as a master student, the potential for less precise scores during the initial stages of data collection is acknowledged, drawing a broad analogy with the learning curve of a neural network. This underscores the importance of experience and exposure to a

diverse range of cases for enhancing accuracy over time.

To end the discussion chapter, DRNNAGE presents moderate accuracy (78.9%) with nuanced performance on the Dutch medieval sample, as reflected in a mean absolute error of 10.423 years, which gives need for a critical evaluation of the NNs application in specific contexts. The influence of age categories emerges as an important factor, overshadowing the effect of sex and preservation. The multifactorial approach by Navega *et al.* (2022), validated by Rizos *et al.* (2023) and this study, emphasizes the diverse effect of anatomical regions, with cranial sutures showing the highest validity and the vertebrae the lowest. In addition, the factors that influence the NNs capabilities further create complexities of age estimation. When comparing DRNNAGE with traditional methods, such as the Suchey and Brooks method, the NNs performance, it is safe to propose that it is slightly superior. To improve the software's accuracy and applicability, the study suggests expanding the training dataset to include various populations.
In the following, and last chapter, the conclusions will be stated regarding the discussion chapter and the study altogether. Suggestions for further research will be mentioned.

Conclusions

The aim of this study is to evaluate the accuracy, repeatability and limitations of the DRNNAGE software in estimating the age-at-death of adult skeletal remains from a Dutch medieval sample. The multifactorial approach imposed by the DRNNAGE software demonstrated reasonable accuracy in predicting age-at-death, predicting the age correctly in 78.9% of the cases. However, a nuanced performance was observed in comparison to the study of Navega *et al.,* (2022), reflected in a mean absolute error of 10.423 years, deviating from the approximately 6 years reported in previous research. Furthermore, this result is not in vain as the model showed improved performance, particularly for individuals over fifty years of age. This finding recommends the DRNNAGE model as a commendable approach for the age-at-death estimation of elderly adults. As the estimation of age-at-death for elderly people proves difficult, it is one of the most important problems that DRNNAGE might be the solution for and where it could really stand out, as the method seems to be robust to the influence of sex and preservation, deeming it applicable in archaeological contexts.

When compared with traditional age estimation methods, such as the Suchey and Brooks method, DRNNAGE revealed similar performance but fell short of the high accuracy rates often claimed by other studies that employed AI (Czibula *et al.*, 2016; Navega *et al.*, 2018, 2022; Toneva *et al.*, 2020). The study demonstrated age category as a significant factor influencing age-at-death estimation. In addition, an interesting pattern emerged, what indicates potential clustering of incorrect estimates raising questions about the impact of the nearest neighbor processing technique and the potential biases introduced by dataset composition. Several limitations were identified, including the influence of sex-specific factors, the complex nature of multifactorial methods, potential bias in the training dataset, and challenges associated with preservation, taphonomic changes, and inter observer error. While DRNNAGE holds promise for age estimation in forensic anthropology and osteology, this study emphasizes the need for careful interpretation and refinement.

Future studies should investigate into factors influencing age-at-death estimation, considering the impact of sex-specific methods, multifactorial approaches, and the potential biases introduced by varying dataset compositions. Addressing the desirable adjustments in the methodology, in this study, especially concerning the vertebrae, can contribute to improved accuracy of the DRNNAGE software. Expanding the training dataset to include more diverse populations is essential as the NN cannot only be trained with European populations but also with populations from other continents. Cross-validation studies across different ethnicities and regions can enhance the generalizability of the method.

In addition, it is often difficult to obtain data that is unbiased if it is obtained by human visual assessment. This potential bias is introduced when humans from different backgrounds make educated guesses based on experience, which is similar to inter observer error, but in a broader sense defined as human bias. In this study, it is the scoring of all skeletal features which are necessary for the implementation of the NN, introducing bias before the data analysis. If this part could also be adopted by AI, it would lower (human) bias considerably. Suggestions are scanning photographic images or videos from features, bones, 3D scans, etc. of skeletons from which AI models can identify the necessary features and obtain scores. Štepanovský *et al.* (2023), has set an example in this field. That

the results do not contribute much yet, the ideas and the will to conduct these ideas is admirable and will bring changes to the osteological and forensic anthropological field eventually. In conclusion, to improve the overall accuracy of age-at-death estimation, the need for visual assessment by a human should be avoided, because of its subjectiveness and the fail to capture the diversity of morphological changes. Focus to objectify the assessment and analysis of the skeletons by AI (Kotěrová *et al.*, 2018, p.173). To make it seem less grim for the future osteologist who enjoy this kind of practice, this reality is still uncertain and cannot be implemented without the interpretation of human experts, as AI should be utilized as a tool to improve the work of such experts and not replace them. It is right to say that all contributing to the field of osteology and forensic anthropology should be aware for future influences of AI.

Abstract

Exploring the junction of artificial intelligence (AI), osteology and, forensic anthropology, this thesis validates the application of neural networks (NN) for accurate age-at-death estimation in skeletal remains. Because of the lack of accurate age-at-death estimation methods and the discrepancy between biological and chronological ages, there a high demand for objective and unbiased approach. A previous study developed DRNNAGE, a NN solution to estimate age-at-death and reported a promising accuracy in predicting age-at-death (95%) and a mean absolute error (MAE) of ~6 years. In this study, the reproducibility and the accuracy of DRNNAGE prediction will be validated employing an archaeological Dutch medieval skeletal sample from the Middenbeemster collection (MB11). The sample consisted of 52 individuals with an age range of 19-101 years. Through a multifactorial transition analysis, 101 features were scored according to two or three levels of senescent change. The results show that the DRNNAGE provides a considerably reliable estimate of age-at-death with an accuracy of 87.9%, with a relatively strong correlation between the estimated and archival ages (Spearman's r=0.7204). Interestingly, DRNNAGE performed with improved accuracy on individuals over 50 years. In conclusion, DRNNAGE is recommended for applications in elderly individuals and is suitable in archaeological contexts. Further research into different population contexts is needed. As the implementation of AI is still in the early stages, the possibilities of AI collaboration can achieve are infinite.

References

Agarwal, S. C., & Grynpas, M. D. (2009). Measuring and interpreting age-related loss of vertebral bone mineral density in a medieval population. *American Journal of Physical Anthropology*, 139(2), 244-252. https://doi.org/10.1002/ajpa.20977

Aggarwal, C. C. (2018). Neural Networks and Deep Learning: A Textbook. Springer International Publishing. https://doi.org/10.1007/978-3-319-94463-0

Albert, M., Mulhern, D., Torpey, M. A., & Boone, E. (2010). Age estimation using thoracic and first two lumbar vertebral ring Epiphyseal union. *Journal of Forensic Sciences*, 55(2), 287-294. https://doi.org/10.1111/j.1556-4029.2009.01307.x

Bailey, C., & Vidoli, G. (2023). Age-at-Death estimation: Accuracy and reliability of common age-reporting strategies in forensic anthropology. *Forensic Sciences*, 3(1), 179-191. https://doi.org/10.3390/forensicsci3010014

Bartelink, N. V. (2019). *Forensic Anthropology: current methods and practice*. Elsevier Academic Press.

Bedford, M. E., Russell, K. F., Lovejoy, C. O., Meindl, R. S., Simpson, S. W., & Stuart-Macadam, P. L. (1993). Test of the multifactorial aging method using skeletons with known ages-at-death from the grant collection. *American Journal of Physical Anthropology*, *91*(3), 287-297. https://doi.org/10.1002/ajpa.1330910304

Bewes, J., Low, A., Morphett, A., Pate, F. D., & Henneberg, M. (2019). Artificial intelligence for sex determination of skeletal remains: Application of a deep learning artificial neural network to human skulls. *Journal of Forensic and Legal Medici*ne, 62, 40–43. https://doi.org/10.1016/j.jflm.2019.01.004

Blau, S., Ubelaker, D. H., & World. (2009). *Handbook of forensic anthropology and archaeology*. Left Coast.

Boldsen, J. L., Milner, G. R., Konigsberg, L. W., & Wood, J. W. (2002). Transition analysis: A new method for estimating age from skeletons. *Paleodemography*, 73106. https://doi.org/10.1017/cbo9780511542428.005

Buckberry, J., & Chamberlain, A. (2002). Age estimation from the auricular surface of the Ilium: A revised method. *American Journal of Physical Anthropology*, 119(3), 231-239. https://doi.org/10.1002/ajpa.10130

Buikstra, J.E., Ubelaker, D.H., (1994). *Standards for Data Collection from Human Skeletal Remains*. Arkansas Archeological Survey Research Series, 44, Arkansas Archeological Survey, Fayetteville.

Buk, Z., Kordik, P., Bruzek, J., Schmitt, A., & Snorek, M. (2012). The age at death assessment in a multi-ethnic sample of pelvic bones using nature-inspired data mining methods. *Forensic Science International*, 220(1–3), 294.e1-294.e9. https://doi.org/10.1016/j.forsciint.2012.02.019

Calce, S. E. (2012). A new method to estimate adult age-at-death using the acetabulum. American *Journal of Physical Anthropology*, 148(1), 11-23. https://doi.org/10.1002/ajpa.22026

Christensen, A. M., Passalacqua, N. V., & Bartelink, E. J. (2014). *Forensic anthropology: current methods and practice*. Academic Press.

Corsini, M.-M., Schmitt, A., & Bruzek, J. (2005). Aging process variability on the human skeleton: Artificial network as an appropriate tool for age at death assessment. *Forensic Science International*, 148(2–3), 163–167. https://doi.org/10.1016/j.forsciint.2004.05.008

Czibula, G., Ionescu, V.-S., Miholca, D.-L., & Mircea, I.-G. (2016). Machine learning-based approaches for predicting stature from archaeological skeletal remains using long bone lengths. *Journal of Archaeological Science*, 69, 85–99. https://doi.org/10.1016/j.jas.2016.04.004

DiGangi, E. A., Bethard, J. D., Kimmerle, E. H., & Konigsberg, L. W. (2009). A new method for estimating age-at-death from the first rib. *American Journal of Physical Anthropology*, 138(2), 164-176. https://doi.org/10.1002/ajpa.20916

Dirkmaat, D. (2012). A companion to forensic anthropology. Wiley-Blackwell.

Falys, C. G., & Prangle, D. (2014). Estimating age of mature adults from the degeneration of the sternal end of the clavicle. *American Journal of Physical Anthropology*, 156(2), 203-214. https://doi.org/10.1002/ajpa.22639

Front Matter. (2020). *In Statistics and Probability in Forensic Anthropology* (pp. i–ii). Elsevier. https://doi.org/10.1016/B978-0-12-815764-0.09995-0

Gurney, K. N. (1997, January 1). *An Introduction to Neural Networks*.

Henderson, C. Y., Mariotti, V., Pany-Kucera, D., Villotte, S., & Wilczak, C. (2012). Recording specific Entheseal changes of Fibrocartilaginous Entheses: Initial tests using the Coimbra method. *International Journal of Osteoarchaeology*, 23(2), 152-162. https://doi.org/10.1002/oa.2287

Hoppa, R. D., & Vaupel, J. W. (2008, October 30). *Paleodemography*. Cambridge University Press.

Humphrey, L. T. (1998). Growth patterns in the modern human skeleton. *American Journal of Physical Anthropology*, *105*(1), 57–72. https://doi.org/10.1002/(sici)1096-8644(199801)105:1%3C57::aid-ajpa6%3E3.0.co;2-a

İşcan, M. Y., Loth, S. R., & Wright, R. K. (1984). Metamorphosis at the sternal rib end: A new method to estimate age at death in white males. American Journal of Physical Anthropology, 65(2), 147–156. https://doi.org/10.1002/ajpa.1330650206

Katz, D., & Suchey, J. M. (1986). Age determination of the male Os pubis. American Journal of Physical Anthropology, 69(4), 427–435. https://doi.org/10.1002/ajpa.1330690402

Kotěrová, A., Navega, D., Štepanovský, M., Buk, Z., Brůžek, J., & Cunha, E. (2018). Age estimation of adult human remains from hip bones using advanced methods. *Forensic Science International*, 287, 163–175. https://doi.org/10.1016/j.forsciint.2018.03.047

Latham, K. E., Finnegan, M., & Rhine, S. (2010). *Age estimation of the human skeleton*. Charles C. Thomas.

Loth, S. R., İşcan, M. Y., & Scheuerman, E. (1994). Intercostal variation at the sternal end of the rib. *Forensic Science International*, 65(2), 135-143. https://doi.org/10.1016/0379-0738(94)90268-2

Lovejoy, C. O., Meindl, R. S., Mensforth, R. P., & Barton, T. J. (1985a). Multifactorial determination of skeletal age at death: A method and blind tests of its accuracy. *American Journal of Physical Anthropology*, 68(1), 1–14. https://doi.org/10.1002/ajpa.1330680102

Lovejoy, C. O., Meindl, R. S., Pryzbeck, T. R., & Mensforth, R. P. (1985b). Chronological metamorphosis of the auricular surface of the ilium: A new method for the determination of adult skeletal age at death. *American Journal of Physical Anthropology*, 68(1), 15–28. https://doi.org/10.1002/ajpa.1330680103

Luna, L. H., & Aranda, C. M. (2022). Adult age-at-death estimation using the first rib: A simple probabilistic approach. *Journal of Forensic Sciences*, 67(6), 2173-2191. https://doi.org/10.1111/1556-4029.15119

Martrille, L., Ubelaker, D. H., Cattaneo, C., Seguret, F., Tremblay, M., & Baccino, E. (2007). Comparison of Four Skeletal Methods for the Estimation of Age at Death on White and Black Adults. *Journal of Forensic Sciences*, *52*(2), 302–307. https://doi.org/10.1111/j.1556-4029.2006.00367.x

Meindl, R. S., & Lovejoy, C. O. (1985). Ectocranial suture closure: A revised method for the determination of skeletal age at death based on the lateral-anterior sutures. *American Journal of Physical Anthropology*, 68(1), 57–66. https://doi.org/10.1002/ajpa.1330680106

Milner, G. R., & Boldsen, J. L. (2012). Skeletal age estimation: Where we are and where we should go. *A Companion to Forensic Anthropology*, 224-238. https://doi.org/10.1002/9781118255377.ch11

Navega, D., Coelho, J. d'Oliveira, Cunha, E., & Curate, F. (2018). DXAGE: A New Method for Age at Death Estimation Based on Femoral Bone Mineral Density and Artificial Neural Networks. *Journal of Forensic Sciences*, 63(2), 497–503. https://doi.org/10.1111/1556-4029.13582

Navega, D., Costa, E., & Cunha, E. (2022). Adult Skeletal Age-At-Death Estimation through Deep Random Neural Networks: A New Method and Its Computational Analysis. *Biology*, *11*(4), 532. https://doi.org/10.3390/biology11040532

Navega, D., & Cunha, E. (2020). Extreme learning machine neural networks for adult skeletal age-at-death estimation. *Statistics and Probability in Forensic Anthropology*, 209-225. https://doi.org/10.1016/b978-0-12-815764-0.00019-8

Passalacqua, N. V. (2009). Forensic age-at-Death estimation from the human sacrum. *Journal of Forensic Sciences*, 54(2), 255-262. https://doi.org/10.1111/j.1556-4029.2008.00977.x

Rissech, C., Estabrook, G. F., Cunha, E., & Malgosa, A. (2006). Using the acetabulum to estimate age at death of adult Males. *Journal of Forensic Sciences*, 51(2), 213-229. https://doi.org/10.1111/j.1556-4029.2006.00060.x

Rizos, L., Garoufi, N., Valakos, E., Nikita, E., & Chovalopoulou, M. (2023). Testing the accuracy of the DRNNAGE software for age estimation in a Modern Greek sample. *International Journal of Legal Medicine*. https://doi.org/10.1007/s00414-023-03129-4

Ruengdit, S., Troy Case, D., & Mahakkanukrauh, P. (2020). Cranial suture closure as an age indicator: A review. *Forensic Science International*, *307*, 110111. https://doi.org/10.1016/j.forsciint.2019.110111

San-Millán, M., Rissech, C., & Turbón, D. (2016). New approach to age estimation of male and female adult skeletons based on the morphological characteristics of the acetabulum. International *Journal of Legal Medicine*, 131(2), 501-525. https://doi.org/10.1007/s00414-016-1406-4

Schanandore, J. V., Wolden, M., & Smart, N. (2022). The accuracy and reliability of the Suchey–Brooks pubic symphysis age estimation method: Systematic review and meta-analysis. *Journal of Forensic Sciences*, 67(1), 56–67. https://doi.org/10.1111/1556-4029.14911

Shook, B., Nelson, K., & Aguilera, K. (2019). *Explorations: An open invitation to biological anthropology*.

Sluis, I. F., Bartholdy, B. P., Hoogland, M. L., & Schrader, S. A. (2022). Age estimation using vertebral bone spurs; Testing the efficacy of three methods on a European population. *Forensic Science International: Reports*, 6, 100301. https://doi.org/10.1016/j.fsir.2022.100301

Snodgrass, J.J. (2004). Sex Differences and Aging of the Vertebral Column. *J. Forensic Sci. 49*, 1–6.

Štepanovský, M., Buk, Z., Pilmann Kotěrová, A., Brůžek, J., Bejdová, Š., Techataweewan, N., & Velemínská, J. (2023). Automated age-at-death estimation from 3D surface scans of the facies auricularis of the pelvic bone. *Forensic Science International*, 349, 111765. https://doi.org/10.1016/j.forsciint.2023.111765

Todd, T. W. (1920). Age changes in the pubic bone. I. The male white pubis. *American Journal of Physical Anthropology*, 3(3), 285–334. https://doi.org/10.1002/ajpa.1330030301

Toneva, D., Nikolova, S., Agre, G., Zlatareva, D., Hadjidekov, V., & Lazarov, N. (2020). Machine learning approaches for sex estimation using cranial measurements. *International Journal of Legal Medicine*, 135(3), 951–966. https://doi.org/10.1007/s00414-020-02460-4

Ubelaker, D. H., & Khosrowshahi, H. (2019). Estimation of age in forensic anthropology: Historical perspective and recent methodological advances. *Forensic Sciences Research*, 4(1), 1–9. https://doi.org/10.1080/20961790.2018.1549711

Wang, X., Liu, Y., & Xin, H. (2021). Bond strength prediction of concrete-encased steel structures using hybrid machine learning method. *Structures*, *32*, 2279–2292. https://doi.org/10.1016/j.istruc.2021.04.018

Watanabe, S., & Terazawa, K. (2006). Age estimation from the degree of osteophyte formation of vertebral columns in Japanese. *Legal Medicine*, 8(3), 156-160. https://doi.org/10.1016/j.legalmed.2006.01.001