



8-2018

## Evaluating the Efficacy of Convolution Neural Networks in Age at Death Estimation Using 3D Scans of the Pubic Symphyseal Face

Melissa A. Brown  
*Western Michigan University*

Follow this and additional works at: [https://scholarworks.wmich.edu/masters\\_theses](https://scholarworks.wmich.edu/masters_theses)



Part of the Anthropology Commons

---

### Recommended Citation

Brown, Melissa A., "Evaluating the Efficacy of Convolution Neural Networks in Age at Death Estimation Using 3D Scans of the Pubic Symphyseal Face" (2018). *Masters Theses*. 3695.

[https://scholarworks.wmich.edu/masters\\_theses/3695](https://scholarworks.wmich.edu/masters_theses/3695)

This Masters Thesis-Open Access is brought to you for free and open access by the Graduate College at ScholarWorks at WMU. It has been accepted for inclusion in Masters Theses by an authorized administrator of ScholarWorks at WMU. For more information, please contact [wmu-scholarworks@wmich.edu](mailto:wmu-scholarworks@wmich.edu).



EVALUATING THE EFFICACY OF CONVOLUTION NEURAL NETWORKS  
IN AGE AT DEAT ESTIMATION USING 3D SCANS OF  
THE PUBIC SYMPHYSEAL FACE

by

Melissa A. Brown

A thesis submitted to the Graduate College  
in partial fulfillment of the requirements  
for the degree of Master of Arts  
Anthropology  
Western Michigan University  
August 2018

Thesis Committee:

Jacqueline Eng, Ph.D., Chair  
Michelle Machicek, Ph.D.  
Britt Hartenberger, Ph.D.

© 2018 Melissa A. Brown

# EVALUATING THE EFFICACY OF CONVOLUTION NEURAL NETWORKS IN AGE AT DEATH ESTIMATION USING 3D SCANS OF THE PUBIC SYMPHYSEAL FACE

Melissa A. Brown, M.A.

Western Michigan University, 2018

The research presented assesses the utility of machine learning approaches, specifically convolutional neural networks (CNNs), to the estimation of age at death in adult decedents by analysis of the pubic symphyseal face of the os coxa rendered as a 3D image. Age at death estimation is an important duty of forensic anthropologists working in medico-legal contexts, as well as bioarcheological researchers. The purpose of this study is to evaluate the accuracy of a CNN relative to the performance of human observers using traditional methods of age estimation. To accomplish this, a CNN created for this project and expert anthropologists were tasked to provide age at death estimates for a selected population with a known age at death. CNN estimation is expected to achieve good accuracy among young adults, and to outperform humans among older adults.

A critical evaluation of the challenges associated with integration of machine learning technology to applied forensics is provided, as well as a novel strategy for interpretation of CNN age at death determinations. The results of the research indicate that, contrary to expectation, CNN age at death estimates are not accurate among young or older adults. Human estimates are superior for these age groups. However, CNN results are surprisingly highly accurate among the middle aged. A number of confounding factors, including primarily population biases in the training set, may contribute to these results. A review of future strategies for improvement of CNN performance is offered.

## ACKNOWLEDGMENTS

First and foremost, my deepest thanks and all my gratitude to James and Dillon. I am amazed at all the hard work you put into this project for me! I literally could not have done it without you. I couldn't have asked for better collaborators, and I treasure our philosophy debates as fondly as our on-topic discussions.

I would like to extend my heartfelt appreciation to Dr. Laura Fulginiti for her generosity while hosting me during my data collection, and her advice and assistance throughout and beyond. To my committee - Dr. Jacqueline Eng, Dr. Michelle Machicek, and Dr. Britt Hartenberger, I thank you for all your encouragement, critiques, and deadline latitude! I consider myself fortunate to have such a wonderful team on my side during the process.

To my cats who never stepped on the keyboard until I hit 'save', and all the friends and family who helped along the way, all my love and gratitude.

Melissa A. Brown

## TABLE OF CONTENTS

ACKNOWLEDGMENTS.....	ii
LIST OF TABLES.....	vi
LIST OF FIGURES.....	vii
CHAPTER	
I. INTRODUCTION.....	1
Age at Death Estimation.....	1
Research Objectives.....	2
Introduction to Machine Learning.....	3
Convolutional Neural Networks.....	4
II. METHODS.....	7
Overview of the Convolutional Neural Network.....	7
Population Studied.....	8
Creating the Training and Test Sets.....	8
Scanning and Post-Processing.....	8
Voxelizing Scans for CNN Analysis.....	10
Selecting the Training Set.....	11
Selecting the Test Set.....	12
Inter-observer Classifications.....	17

## Table of Contents—Continued

### CHAPTER

III.	RESULTS.....	20
	Results by Individual.....	20
	Results by Age Category.....	20
	Young Adults.....	20
	Middle Adults.....	22
	Old Adults.....	24
	Overall Results.....	26
	CNN Results.....	26
	Observer Results.....	27
IV.	DISCUSSION.....	31
	Integrating and Interpreting Machine Learning Systems.....	31
	The Black Box Problem.....	31
	Challenges to Integration.....	32
	Strategy for Interpretation.....	35
	Individual Case Studies.....	38
	General Context for Case Studies.....	38
	Individual 01.....	39
	Individual 02.....	42
	Individual 06.....	44

## Table of Contents—Continued

### CHAPTER

Individual Case Studies.....	38
Individual 16.....	47
Broad Considerations.....	51
Age Cohort Classification as a Proxy for Accuracy.....	51
Among Young Adults.....	53
Among Middle Adults.....	54
Among Older Adults.....	57
V. CONCLUSIONS AND FUTURE DIRECTIONS.....	59
Applied Utility.....	59
Project Limitations.....	60
Training Sample Size.....	60
Scanning Technique.....	61
Computational Variables.....	63
FANSIE in the Future.....	64
WORKS CITED.....	67
APPENDICES	
A. Technical Report by James Jenkins.....	69
B. Results by Individual.....	70



## LIST OF TABLES

2.1 Test Set: Restricted Age Categories.....	13
2.2 Test Set: Individual Census.....	13
2.3 Test Set: Individual Representation.....	14
2.4 Final Test Set.....	16
3.1 Young Adult CNN Results.....	21
3.2 Young Adult Observer Scores.....	22
3.3 Middle Adult CNN Results.....	22
3.4 Middle Adult Observer Scores.....	24
3.5 Old Adult CNN Results.....	24
3.6 Old Adult Observer Scores.....	26
3.7 Overall Results: Known Age, CNN, and Observers.....	29

## LIST OF FIGURES

1.1 CNN Operations: Input.....	5
1.2 CNN Operations: Activation.....	5
2.1 Scan Editing.....	9
2.2 Meshed and Single Point Vertices.....	10
2.3 Boxed Voxelized Pubic Symphysis Scan.....	11
2.4 Suchey-Brooks Phases.....	18
3.1 Young Adult CNN Results.....	21
3.2 Middle Adult CNN Results.....	23
3.3 Old Adults CNN Results.....	25
3.4 Overall Results: Known Age, CNN, and Observer 1.....	30
4.1 Individual 01: Known Age, CNN Age, and Observer Scores.....	39
4.2 Individual 01: Images of Pubic Symphyseal Face.....	40
4.3 Individual 01: Voxelization of 3D Scan.....	41
4.4 Representative Images from Training Set Displaying Remnant Billowing.....	41
4.5 Individual 02: Known Age, CNN Age, and Observer Scores.....	43
4.6 Individual 02: Images of Pubic Symphyseal Face.....	43
4.7 Individual 06: Known Age, CNN Age, Observer Scores.....	45
4.8 Individual 06: Images of the Pubic Symphyseal Face.....	45
4.9 Individual 06: Voxelization of 3D Scan.....	46

## List of Figures—Continued

4.10 Individual 16: Known Age, CNN Age, Observer Scores.....	48
4.11 Individual 16: FSC Method Age Estimation Results.....	49
4.12 Individual 16: Images of Pubic Symphyseal Face.....	50
4.13 Representative Images from the Training Set, Geriatrics.....	50
4.14 Individual 15: Images of Pubic Symphyseal Face.....	51
4.15 CNN Age Output Trends.....	55
5.1 High Def and Low Def Scan Techniques.....	62

## CHAPTER I

### INTRODUCTION

#### Age at Death Estimation

One of the essential duties of anthropologists of human biology involves the construction of the biological profile, which refers to a compendium of evidentiary supported demographic data that is typically derived from osteological analysis. Skeletal material is well suited to this goal, and can be particularly useful in forensic or archaeological assessments because bones preserve more readily than other tissues, while still bearing substantial information about the individual. Information gained includes estimation of a number of useful characteristics, such as sex, ancestry, and age at death. These qualities are often extrapolated based on assessment of morphological features displaying variable, but predictable, degrees of bony degradation or development of particular features. Estimation of age at death in adult individuals may be achieved by assessment of a number of different osteological phenomena, however the standard method is evaluation of the pubic symphyseal face of the os coxa (Ritz-Timme *et al.*, 2000).

The pubic symphyseal face is the articulation surface of the pubis bone of the os coxa, wherein the left and right aspects of the pelvic girdle connect via fibrocartilage. The surface of the pubic symphysis undergoes relatively predictable morphological transformations throughout life that are associated with ageing. During adolescence and into early adulthood, the pubic symphysis bears a distinct morphological patterning. However, throughout adulthood, bony transformations ensue which alter surface morphology. Pubic symphysis transformations occur in predictable patterns that are correlated with progressing age. A number of strategies have been developed to classify these changes, and therefore estimate the age of adult individuals (see most notably Todd, 1920; Brooks and Suchey, 1990; and Hartnett, 2010, among many others).

The methods above rely upon the subjective assessment of morphological features by the osteologist, and therefore such strategies may lead to flaws in age estimation. Underestimation of

age in elderly individuals is a well-known problem in the discipline. The overlapping age ranges that these phase scoring methods produce also tend to yield very broad estimations, which may be at odds with other age estimation techniques (Milner and Boldsen, 2012). Furthermore, research indicates that the discipline's standard, the Suchey-Brooks method, is not an extremely accurate technique for determining age at death based on skeletal morphology (Hartnett, 2010). These problematic aspects of age estimation are particularly important to address in forensic applications where the capacity to produce accurate age estimations has important legal and social implications.

### Research Objectives

Increasingly, research conducted on pubic symphysis ageing methods has emphasized quantitative or computational approaches that aim to reduce subjective bias and provide more discrete age categorization (see Biwasaka, 2013; Slice and Algee-Hewitt, 2015; Stoyanova *et al.*, 2015). This project furthers that trend by assessing the utility of a machine learning based approach to age at death estimation, specifically the use of a convolutional neural network, or 'CNN', created to evaluate 3D digital renders of the pubic symphysis. Machine Learning technology is rapidly being applied to a number of novel research contexts, and with admirable success. At the time of this writing, however, machine learning is only tentatively being considered for use in anthropology.

The research presented in this thesis represents an important contribution to the anthropological sciences by evaluating the utility of machine learning technology in the context of human skeletal analysis. Preliminary studies such as this, which aim to establish the feasibility of machine learning approaches to age at death estimation, are an important first step to determining if these technological approaches will be of greater use to osteologists than current methods, or may function as useful tools to augment traditional methods of data collection. Therefore, the research presented attempts to assess if CNNs are able to produce an accurate age at death estimation, and if so if that performance is superior to that of human osteologists using standard techniques. It is expected that the CNN will be able to produce its most accurate estimations among young adults, who have more morphologically distinct features than other ages. Further, it is expected that the CNN will outperform human devised methods of age

estimation among elderly individuals. In addition to the empirical assessment of a CNN's competency at age-at-death estimation provided herein, this project also presents a novel examination of the potential legal and institutional challenges associated with incorporating machine learning to medico-legal contexts.

## Introduction to Machine Learning

Machine learning is a specialized subfield of computer science that involves the construction of computational algorithms that allow the computer to learn, or make associations among data, without being specifically programmed to do so. There are a variety of different machine learning approaches that are suited to different chores. Artificial neural networks are a special type of machine learning system that are so-called because their construction mimics the neural connectivity of biological organisms.

These systems are comprised of a series of interconnected layers whose operations are contingent upon data received from the preceding layer, in much the same way that a mammalian brain receives sensory input, and then interprets it accordingly based on the qualities of the input data. Neural networks function analogously to biological systems in that the activation of a neuronal cluster, or node, in one layer of the network triggers the subsequent response in the deeper layers before ultimately culminating into some output, the system's thoughts. This is procedurally akin to the way that the activation of a biological neuron or node, in response to the input of sensory data, instigates a chain reaction that eventually characterizes that input, the organism's thoughts.

The manner, and accuracy, with which a neural network characterizes input is dependent upon its experience with similar sorts of data. Just as a human can readily distinguish a dog from a cat based on a lifetime of familiarity with these common companions, a network that has had the opportunity to gain comparable experience through evaluation of identified images of each can be easily be trained to do the same. Where neural networks may have some advantage over humans in performing these kinds of tasks is that networks may assess a tremendous volume of data very quickly, and unlike humans are not prone to subjective opinions. While the actual characterization output of a network may lack accuracy, its decision making will still be inherently predicated upon a mathematically derived system of evaluation. A neural network

may make a poor judgement because of poor training, but cannot make a bad judgment because of a bad day.

## Convolutional Neural Networks

CNNs are machine learning systems that are especially well suited to the evaluation of image files, including 3D images. CNNs are currently being heavily researched for potential utility in computer assisted medical diagnostics, among many other applications, and have demonstrated great success. What gives CNNs their tremendous predictive power in image classification is their ability to recognize normative focal patterns, or localized structures, within the data that are diagnostically meaningful. Such regions are recognized as discrete features, rather than as immutable aspects of the whole. As such, when assessing novel input, the CNN is able to recognize key areas that are important in classification, rather than comparing the whole of one image against another. The process of feature recognition allows for CNNs to correctly classify images with variable, but predictable, relationships among features. The image classification process is illustrated in the figures below using the classic example of handwriting analysis, which prior to the advent of machine learning was a very challenging task for computers.

In this simple example, suppose that a CNN were tasked to correctly identify a written digit. Traditional computational methods of performing such a task would involve comparing the novel input, some handwriting style, against a database of hundreds to hundreds of millions of other examples of that digit. The problem with this type of analysis is that computers think very literally, so unless one of the comparative exemplars in the database is a perfect fit, it probably won't be recognized. CNNs avoid this trap by assessing the image in small units that create an accurate impression of the gestalt, but can reckon with inputs that are larger, smaller, slanted, misshapen or otherwise not exemplary. The flexibility of CNNs is why this approach was favored over other, simpler, machine learning techniques. There is no 'ideal' pubic symphyseal face and other computational approaches can be expected to get misled by the minor morphological differences that are normal expressions of human biological variation.

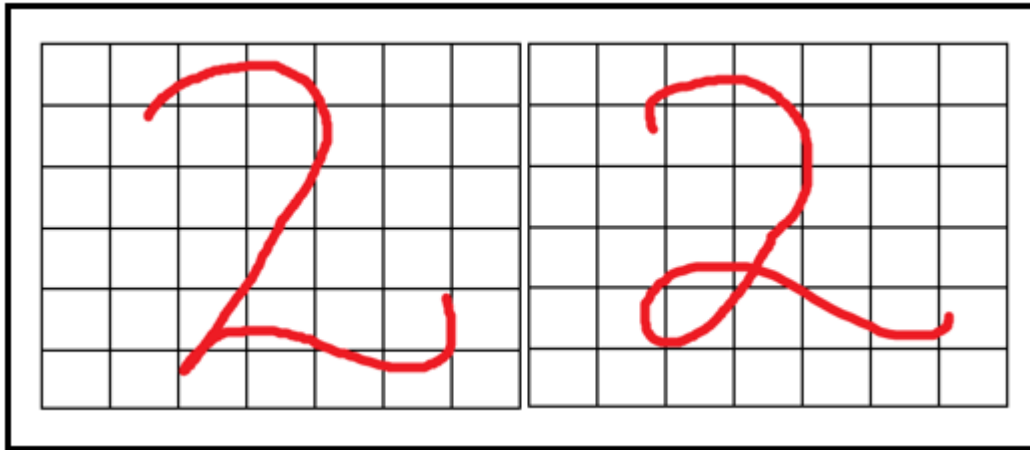


Figure 1.1 CNN Operations: Input

Here, we see two different examples of the number 2 that have been input for analysis. The images have been broken into smaller units for assessment by aligning a grid on top of the number. Each cell of the grid can be likened to a neuron in that it may be active, or bear data within it, or inactive and having no data. Active regions are expressed mathematically with a positive number 1, while inactive regions are assigned a value of -1.

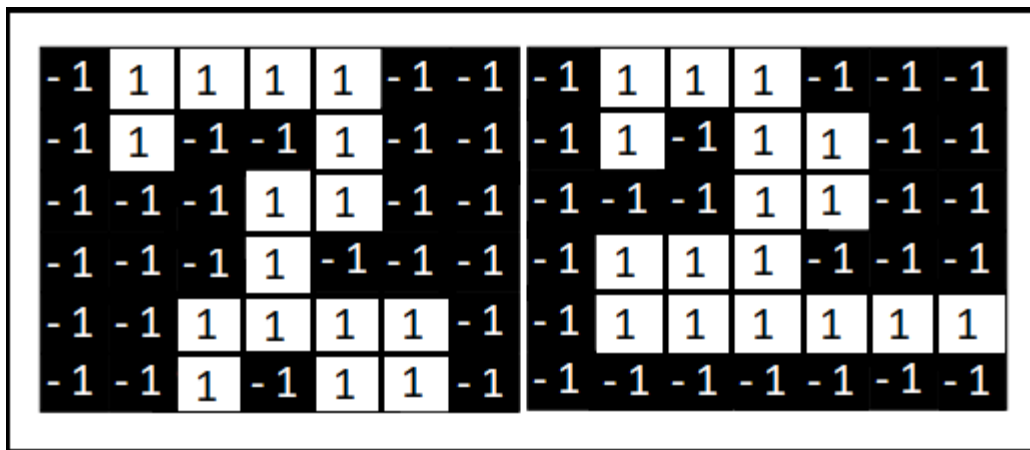


Figure 1.2 CNN Operations: Activation

The image can now also be expressed mathematically as matrix of values. The first written number two can be represented thus:



$$\left\{ \begin{array}{cccccccc} -1 & +1 & +1 & +1 & +1 & -1 & -1 \\ -1 & +1 & -1 & -1 & +1 & -1 & -1 \\ -1 & -1 & -1 & +1 & +1 & -1 & -1 \\ -1 & -1 & -1 & +1 & -1 & -1 & -1 \\ -1 & -1 & +1 & +1 & +1 & +1 & -1 \\ -1 & -1 & +1 & -1 & +1 & +1 & -1 \end{array} \right\}$$

The fundamental nature of the written digit is still retained by this transformation, but the CNN can now recognize features within the data that are important. The CNN teaches itself which features are relevant to classification through its training process, wherein it learns useful associations through the analysis of a known data set. The CNN created for this project was trained on a batch of 3D images of the pubic symphyses from individuals with a known age at death. The intention is that the CNN will learn features that are well suited to producing an age classification of its own accord, and then be able to apply its understanding of those features to novel inputs and achieve accurate results in classification.

The CNN tests each feature it has devised against every aspect of a novel input, and the values of the initial matrix are increased or decreased by simple arithmetic according to how well they fit with diagnostic features. The term for this process is 'convolving', which produces a new matrix of values that mathematically represents how well the novel input matches against known features. The newly created values are known as weights.

Heavily weighted regions, which may be either the initial cell 'neuron', or a collection of neurons in a node, can be thought of as highly active regions, continuing the similarities between CNNs and biological neural processing. The matrix of weights is then progressed, or filtered, through additional layers. Other deep, hidden layers of the CNN work in concert with convolving layers to further transform the input data into an increasingly meaningful matrix of weights that eventually yields an output classification. The research presented will determine if this computational process is capable of generating accurate age-at-death estimations.

## CHAPTER II

### METHODS

#### Overview of the Convolutional Neural Network

The CNN designed for this project was built by collaborators Dillon Daubert and James Jenkins, of Western Michigan University's Computer Sciences Department. The base architecture for the CNN was developed using the Tensorflow™ version 1.7 open source software library. Like many complex applications, the CNN for this project was given a distinct name. Our CNN was designated 'FANSIE', the Forensic Anthropology Neural System Interdisciplinary Experiment, and will be referred to as such throughout this thesis. FANSIE is designed with three different models. Each model filters data through 20 hidden layers. During training FANSIE was provided with the 3D image and the known age-at-death of all individuals in the training set so that associations between age and morphology could be learned.

Validation testing, in accordance to CNN best practices, was conducted throughout training on each model. This is a process to avoid overfitting the CNN, or training it too specifically to the training set to learn features that may be generalizable to new inputs. Twenty percent of the training set was reserved for validation testing, and the individuals in the validation set were chosen at random for each model of the CNN. Upon completion of training and validation, FANSIE was provided a test set of 16 individuals to evaluate (see 'Selecting the Test Set' below). Age determinations were generated at final output using a linear regression function to assign a real number between 0 - 100. The mean of the output from each of the three models was used as the final age determination. Please see Appendix A: CNN Technical Report by James Jenkins for a detailed review of the technical configuration of the CNN.

## Population Studied

### The Hartnett-Fulginiti Collection

The Hartnett-Fulginiti Collection (HFC) is a large (n=620) assemblage of bilateral pubic symphyses and 4th rib sternal portions curated at the Forensic Science Center (FSC) of Maricopa County, Arizona. Specimens are from a modern American population of decedents. Both males and females are present in the collection, and the age of individuals included ranges from 18 - 99 years. Information concerning demography (race, sex, and age) as well as aspects of the decedents' medical history are known. The left and right aspects of the pubic symphyses curated at the FSC were excised from the innominate and surrounding soft tissues by cutting through the superior and inferior pubic rami, and then macerated and dried (Hartnett, 2010). The excision of the pubic symphysis from the whole of the os coxae enables much more effective scanning, as only this feature is relevant to the project. In addition to the large sample size available, these aspects of the HFC make it ideal for this research.

## Creating the Training and Test Sets

### Scanning and Post-Processing

3D scans were generated using a NextEngine 3D Desktop Laser Scanner. A scanning parameter protocol was established which maximized the highest definition capacity of the NextEngine device. The scanning protocol included 16 division scans, 40,000 point per square inch density, bright light settings, and macro focus. Scans were taken using the 'bracketing' approach. The technique captures three views of the pubic symphyseal face - a front facing view, a view with moderate dorsal side rotation, and a view with moderate ventral side rotation.

All available individuals in the HFC were scanned. Generally, only the left aspect of the pubic symphysis was scanned. In circumstances where the left aspect was unavailable (either missing, damaged, or too fragile to mount for scanning) the right was scanned instead. After scanning, individuals were assigned an alphanumeric code denoting the specimen number (adopted from the coding used in the Collection curation) and age-at-death. Modifiers were added to relevant scans to indicate if the scan was right-sided or from a female (e.g. '001-18-f-r'). Photographs were taken of each specimen following scanning and assigned an identical code.

After the initial digitization of the symphysis, individual scans were processed using ScanStudio™ software. Scans were edited to remove extemporaneous data. The scanning platform and various devices used to hold the bones stable were trimmed from each file. Bony scan data not considered contextually relevant for the CNN was also trimmed. No other editing, such as data smoothing or patching, was performed on the scans. Figure 2.1 shows the progression of scan editing.

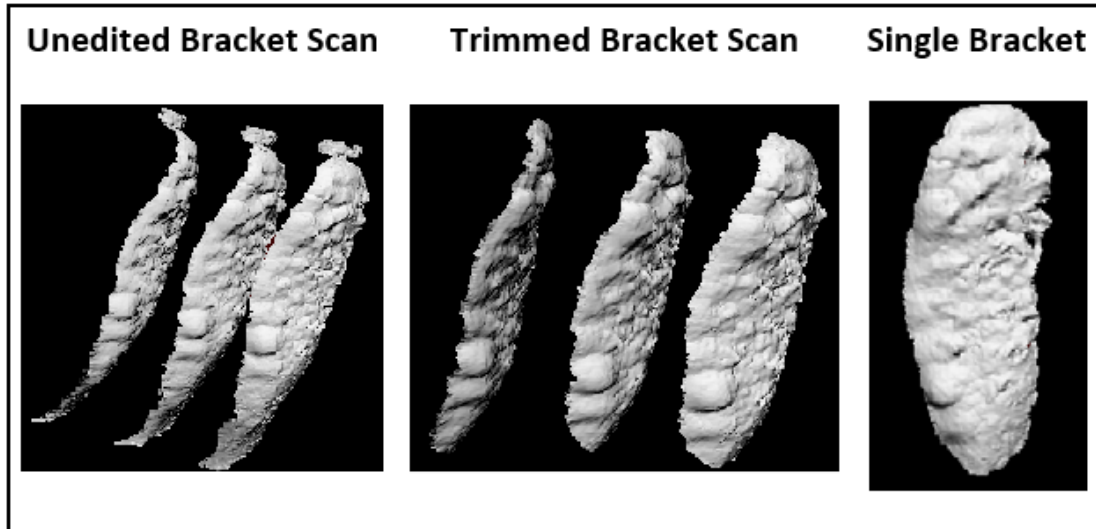


Figure 2.1 Scan Editing

Rather than merge the three aspects of the bracket scan into a single 3D image via pixel matching, they were assigned their own serial code and treated discretely, for a number of reasons. First, it provided a means of augmenting the data set. By treating each aspect of the bracket scan separately, the CNN was given a more robust sample of images to analyze. Further, the minute variations that each aspect bears may be presumed to bolster the CNN's sensitivity to deviations from a hypothetical norm. Finally, due to the shape of the pubic bone, the laser scanner was variably effective at rendering the ventral and dorsal sides of the bone. Because laser light travels in a straight line, while the rami of the pubic bone are contoured, often the lasers would miss the immediate bony material on the sides of the face. Because of this complication, the dorsal and ventral side scans often failed to pick up additional data that the front-facing scan lacked. Regardless, the simple addition of a scan with moderate rotation is itself useful. Comparable strategies of data augmentation, wherein different displays of the same

image are provided to the CNN (e.g. flipped, rotated, or reversed) have been demonstrated to improve classification accuracy (Dieleman *et al.*, 2015). Therefore, regardless of the content of the data itself, all three aspects of the bracket scan were included in the training set. After clean-up, each bracket scan was exported as an individual .obj file and assigned a new code to reflect front view, dorsal view, or ventral view (e.g. 001-18-a.obj, 001-18-b.obj, 001-18-c.obj).

### Voxelizing Scans for CNN Analysis

The .obj file type allows for 3D data to be represented in a way that may be translated into a format that is suitable for CNN analysis. Upon scanning, the 3D topography of the symphyseal face is digitally rendered as a series of adjacent triangles that create a meshed landscape which forms the image. The point of each triangle, wherein it converges with another triangle, is known as the vertex. These vertices are in truth the digital landmark points that the laser scanner generates when impacting with a solid surface, and the adjacent triangles are formed by the software drawing lines between vertices. In this way, each vertex represents a point of spatial orientation within the context of the whole. Figure 2.2 shows a close-up of the PS face scan used in Figure 2.1. The image on the left shows each vertex joined together to create the blanketed mesh of adjacent triangles, and the image on the right shows the same view of the face represented by the points of the vertices.

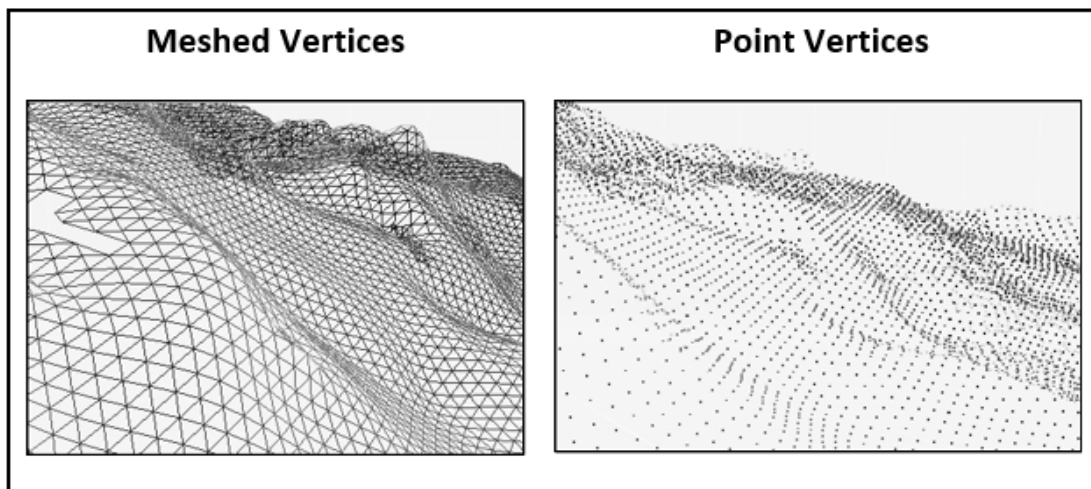


Figure 2.2 Meshed and Single Point Vertices

The spatial relationship among the vertices is used as the foundation for rendering the image as a mass of voxels, or volumized pixels. This strategy enables the diverging and nonhomogeneous topographies of each individual's PS face to be represented in a way that can be analyzed quantifiably, and therefore computationally. To allow the CNN to extrapolate meaning from the scans, voxelized images were bound in a 3D box, with 38x38x38 voxel dimensions. Voxel density, a fair proxy for image resolution, was dictated by the computational power available for training the CNN. Once the voxelized pubic symphysis (PS) scan is boxed, each voxel within functions as an 'active neuron', or meaningful data point, while empty spaces may be regarded as 'inactive neurons'. The overall matrix of active and inactive regions within the boundaries of the box, once the image has been filtered and transformed through the layers of the neural network, is then used as the litmus against which unknown samples are compared to determine a 'best fit' output, or age determination (see Figure 2.3 for a representation of a boxed voxelized scan).

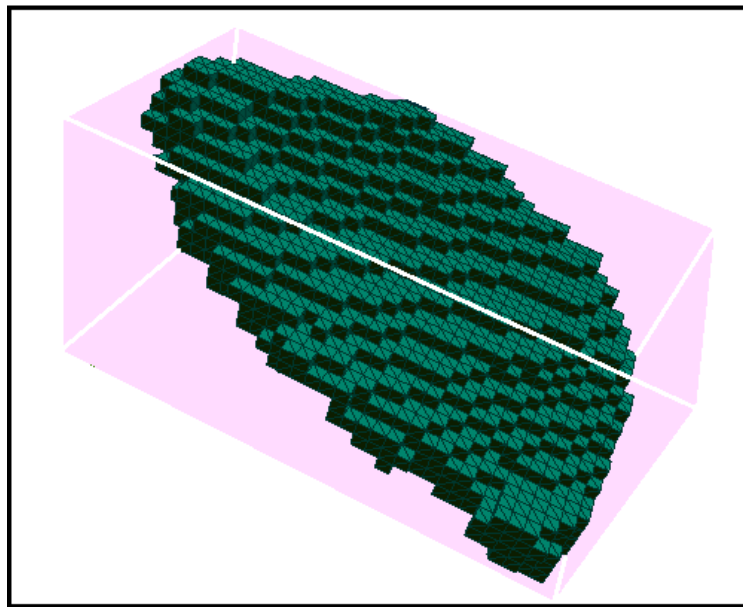


Figure 2.3 Boxed Voxelized Pubic Symphysis Scan

### Selecting the Training Set

To avoid distracting the CNN with any potentially sex-related morphology markers only males were provided for training. Additionally, right-sided scans were eliminated from the

training sample to prevent sidedness from possibly confusing morphology associations. Race was not considered, and individuals of all racial classifications (as assigned by the FSC) were included. Elimination of females and right-sided symphyses yielded a total sample size of  $N=385$ . A sample of  $n=16$  was set aside for final output testing, and not included in the training sample. A validation testing sample was randomly culled from the training set for each model evaluated ( $n=77$ ) and was also excluded from the training sample. This left a final training set size of  $n=292$ .

### Selecting the Test Set

Individuals in the test set were chosen to proportionally represent the distribution of ages in the training sample. The strategy of proportional representation was adopted so that the CNN would be most frequently tested upon the material for which it was best trained. Additionally, selection parameters for individuals in the test set attempted limitation of phase overlap, as defined by the 95% Confidence Interval (CI) of the Suchey-Brooks method phases. Categories were defined such that representative individuals would fall within the spectrum of as few phases as possible, given the broadly overlapping age ranges. For instance, because Phase 4 falls entirely within the spectrum of Phases 3 and 5 no individual from that Phase only could be selected. Further, this method was devised as a means of providing some moderate control when conducting inter-observer scoring tests with volunteer osteologists (refer to the 'Inter-observer Classifications' section for more details). To achieve this, restricted categories were delineated by repurposing the lower and upper thresholds of the 95% CIs of the Suchey-Brooks phases, arranged in ascending order.

Next, all individuals in the overall final sample set were sorted according to their age, and a census was performed to reckon the total number of individuals in each restricted category. Table 2.2 shows the results of the census. Note that the total number of individuals involved in the census is  $n=404$ . The 19 individuals from the census tally absent in the final sample set ( $n=385$ ) are due to missing or inappropriate scans (such as right-sided) not identified until after the test set was isolated.

Table 2.1 Test Set: Restricted Age Categories

Suchey-Brooks Method Age Ranges (95% CI)											
Phase 1		Phase 2		Phase 3		Phase 4		Phase 5		Phase 6	
15 - 23		19 - 34		21 - 46		23 - 57		27 - 66		34 - 86	
Restricted Categories											
15 - 19	20 - 21	22 - 23	24 - 27	28 - 34	35 - 46	47 - 57	58 - 66	67 - 86	87 +		

Table 2.2 Test Set: Individual Census

<b>Restricted Category</b>	<b>N</b>
15 - 19	9
20 - 21	11
22 - 23	10
24 - 27	22
28 - 34	24
35 - 46	82
47 - 57	90
58 - 66	59
67 - 86	83
87+	14

Once the census tally was conducted, the percentage of the total that each category represented was calculated and the number of individuals from each category to be represented in the test set was determined, with the intention of a final test set of 15 individuals. This yielded 14 individuals for the test set, but left the early categories unrepresented due to the low prevalence of individuals from these very narrowly defined ranges, and the general sample set bias towards middle aged and older adults. Table 2.3 displays the results of these calculations, which were performed algorithmically in Python.



Table 2.3 Test Set: Individual Representation

Restricted Category	N (by category)	Percent of Total	N (representative)
15 - 19	9	2.2	0
20 - 21	11	2.7	0
22 - 23	10	2.5	0
24 - 27	22	5.4	1
28 - 34	24	5.9	1
35 - 46	82	20.3	3
47 - 57	90	22.3	3
58 - 66	59	14.6	2
67 - 86	83	20.5	3
87+	14	3.5	1

The early categories were deemed important to include in the test set as they isolated very young adults from falling within the wide margins of the higher phases. Further, due to a random selection process, broadening the parameters of the categories could potentially yield no very young adults at all for the test set. Therefore, the representation strategy was modified for their inclusion. One representative individual was borrowed from the 35 - 46 year old age category. This category, which proportionally has three representative individuals, was reduced to two representatives and the remaining was levied to the empty 15 - 19 year category. Two individuals were added to the test set to provide the 20 - 21 and 22 - 23 age categories with representatives. A final test set number of 16 individuals was thus created, and those individuals were randomly chosen from among their respective categories. These individuals were then recoded by arranging them chronologically and numbered simply 01 - 16 to thus be identified as Individual 01, Individual 02, and so on (abbreviated as I01 - I16 throughout this manuscript).

The boundaries of the restricted categories naturally lent themselves to defining broader age cohorts. Subsequently, the individuals in the test set were further classified as belonging to a young adult (YA), middle aged adult (MA), or older adult (OA) cohort. The YA cohort was defined as individuals aged between 18 - 34 years, and includes I01 - I05. The MA cohort was defined as individuals aged between 35 - 66 years, and includes I06 - I12. The OA cohort was

defined as ages 67 and older, and includes individuals I13 - I16. Sorting individuals into mutually exclusive categories provides a mechanism for gauging relative CNN accuracy, as output is a discrete number rather than the ranged classifications that are characteristic of traditional ageing methods. Using the defined cohorts, FANSIE may be regarded as correct or incorrect if the age determination produced falls within the range for that cohort. Table 2.4 displays the final test set with new identification codes and categorization.

Table 2.4 Final Test Set

Individual	Known Age	Restricted Category	SB Phase	Cohort
01	18	15 - 19	1	Young Adult (18 - 34)
02	21	20 - 21	1, 2, 3	
03	22	22 - 23	1, 2, 3	
04	27	24 - 27	2, 3, 4, 5	
05	28	28 - 34	2, 3, 4, 5	
06	39	35 - 46	3, 4, 5, 6	Middle Adult (35 - 66)
07	40	35 - 46	3, 4, 5, 6	
08	47	47 - 57	4, 5, 6	
09	50	47 - 57	4, 5, 6	
10	57	47 - 57	4, 5, 6	
11	59	58 - 66	5, 6	
12	62	58 - 66	5, 6	
13	68	67 - 86	6	Old Adult (67 - 100)
14	68	67 - 86	6	
15	82	67 - 86	6	
16	91	87 +	none	

## Inter-observer Classifications

To evaluate the competency of the CNN at producing an age determination relative to human interpretation, three volunteer osteologists were recruited for inter-observer trials. The volunteers represented varying experience levels, but were all considered skilled osteologists and included a Diplomate of the American Board of Forensic Anthropology, a PhD, and a PhD candidate. Volunteer observers were able to evaluate the real bone specimen on site at the FSC, rather than review digital scans.

Inter-observer trials were conducted remotely. Observers (may henceforth be known as Obs 1, Obs 2, or Obs 3) were provided with a simple recording form which listed the HFC code for specimens in the test set arranged in ascending numeric order. The individuals in the HFC were coded as they entered the initial study and thus these codes have no bearing on the age of the individual. Therefore, the numbering was considered sufficiently randomized to minimize any potential biasing during evaluations. Observers were asked to categorize individuals in the test set using two methods of age estimation - the Suchey-Brooks method and the FSC method, which was developed from this population. Observers were trained and experienced with both methods of age estimation.

The initial intention of collecting score reports from different age estimation methods was to perform intraclass correlation coefficient observation error statistics and make a study of observer agreement among each other and between the two methods. Ultimately, this was deemed to be beyond scope for the project as the principal topic under investigation is the performance of the CNN. Consequently, the Suchey-Brooks method score reports were used as the foundation for observer interpretation and trends in observer reporting are discussed only as general correlations, rather than as statistically quantified relationships. However, because the FSC method incorporates a later phase 7 category whose range exceeds the limitations of the Suchey-Brooks phase 6, FSC observer reports were used to contextualize discussion of Individual 16. Figure 2.4 displays the Suchey-Brooks phases with abridged phase morphology descriptions and representative photographs to provide context for Observer scoring and discussion of results; refer to Brooks and Suchey's 1990 publication for full text and original illustrations.

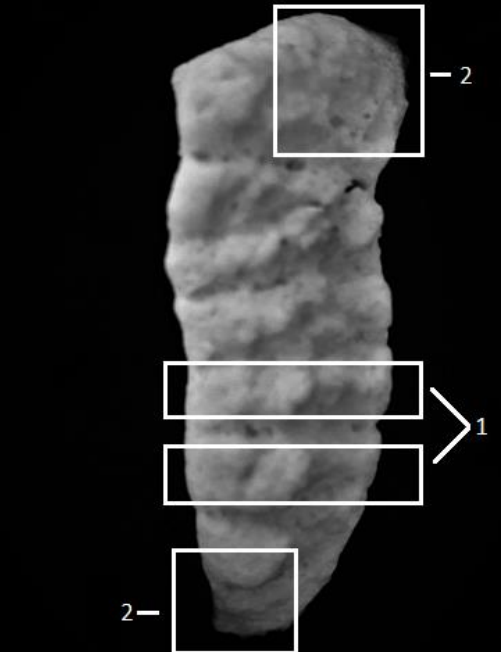
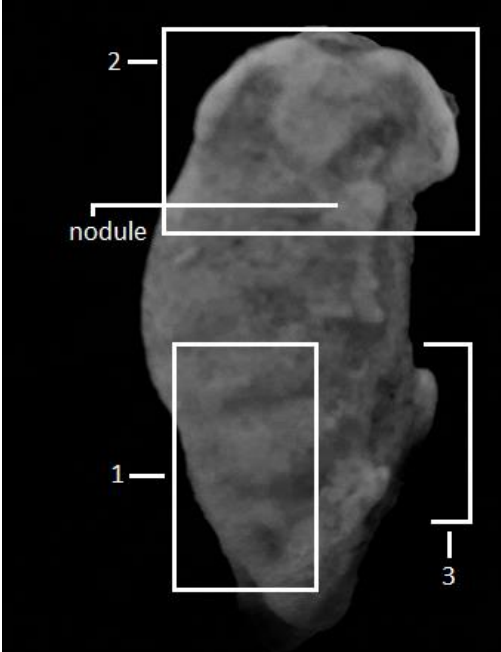
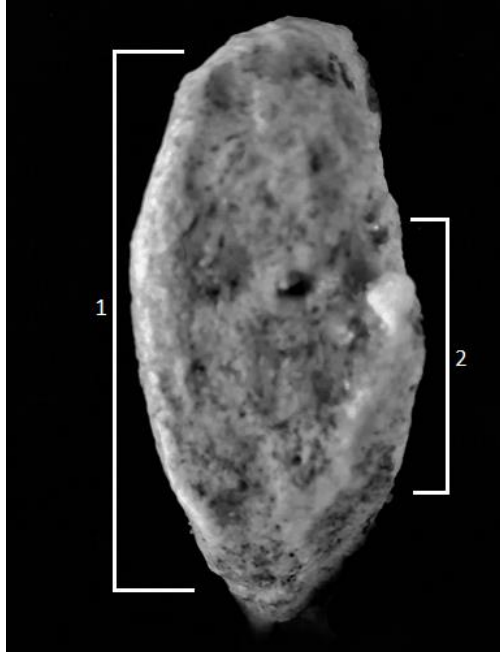
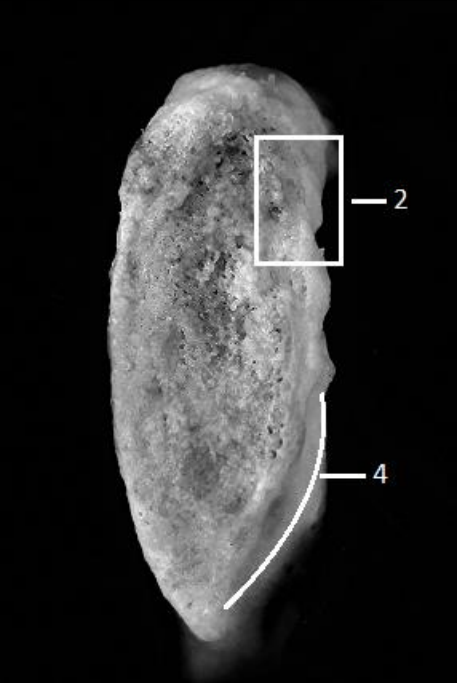
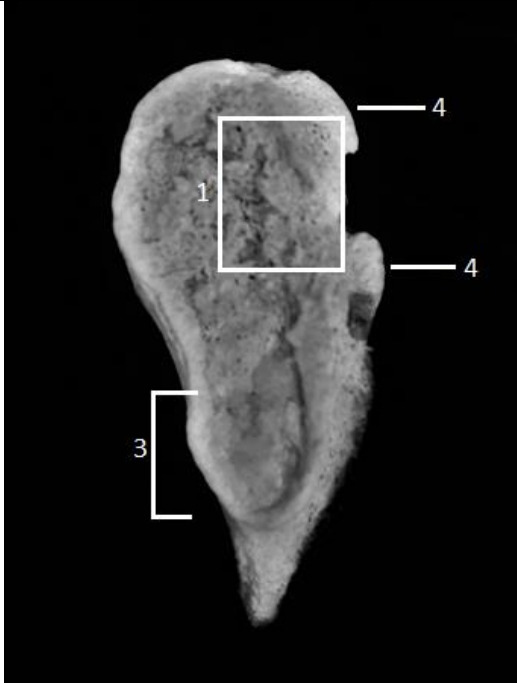
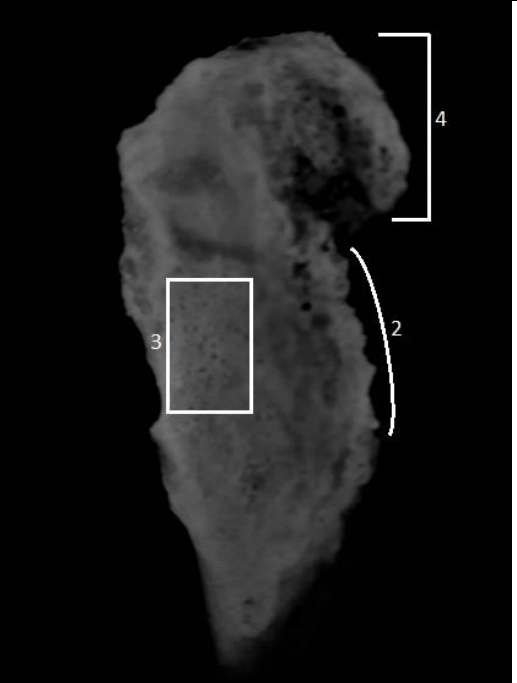
Phase 1 (15 - 23)	Phase 2 (19 - 34)	Phase 3 (21 - 46)
1. System of deep, transverse billowing ridges throughout 2. lack of delimitation of upper or lower extremity (face is contiguous to surrounding bone)	1. ridging may still be apparent 2. early delimitation of upper and lower extremity, perhaps with ossific nodules 3. emerging ventral rampart	1. Complete dorsal plateau 2. ventral rampart near completion 3. face smooth or remnant ridging 4. absence of dorsal lipping, bony growths
		

Figure 2.4 Suchey-Brooks Phases

Figure 2.4 Suchey-Brooks Phases (continued)

Phase 4 (23 - 57)	Phase 5 (27 - 66)	Phase 6 (34 - 86)
<ol style="list-style-type: none"> <li>1. oval outline is complete</li> <li>2. possible hiatus in upper ventral region</li> <li>3. face is generally fine grained</li> <li>4. face may have distinct rim</li> </ol>	<ol style="list-style-type: none"> <li>1. complete rim, with no or only slight erosion along superior ventral region</li> <li>2. face is depressed relative to rim</li> <li>3. lipping along dorsal border</li> <li>4. growths along ventral border</li> </ol>	<ol style="list-style-type: none"> <li>1. facial depression and rim erosion</li> <li>2. pronounced ventral bony growths</li> <li>3. porosities, ossifications</li> <li>4. crenulations</li> <li>5. irregular face shape</li> </ol>
		

## CHAPTER III

### RESULTS

#### Results by Individual

A complete presentation of CNN results for each individual in the test set, and Observer scores using the Suchey-Brooks method, can be found in Appendix B: Individual Results.

#### Results by Age Category

##### Young Adults

Individuals 01 - 05 are classified as young adults. The YA age cohort was defined as 15 - 34 years of age. Individuals tested in this age category range from 18 - 28 years old at age-at-death. The YA cohort represents 18.7% of FANSIE's training set. The CNN failed to accurately classify any individual within this age cohort. FANSIE overestimated age in all young adult individuals with an overall error rate of +23.2 years (Table 3.1; Figure 3.1).

Overall, Observers accurately assigned SBP scores (Table 3.2). While there was SBP score disagreement among observers regarding Individuals 02 and 04, the remainder was unanimously, and correctly, classified. Individual 02 was the only member of the young adult cohort misclassified. All observers overestimated the age of Individual 02. Individual 02 has a known age-at-death of 21 years of age. Observers 1 and 3 classified this individual as SBP 4 (95% CI 23 - 57); Observer 2 classified this individual as phase 5 (95% CI 27 - 66). Notably, this individual also has the highest error rate among CNN generated ages (+ 31.4).

FANSIE consistently overestimated ages of individuals tested from the YA age cohort, and was inaccurate in all classifications. Observers disagreed on phase scores for Individual 04

and 02, but were otherwise unanimous in phase classification. Individual 02 was overestimated in age by all Observers, and least accurate among ages produced by FANSIE.

Table 3.1 Young Adult CNN Results

Young Adult				
Age Range		Training Set %		Mean Error
15 - 34		18.7		+ 23.2
Error Rate by Individual				
Individual		Known Age	CNN Age	CNN Error
01		18	34.2	+ 16.2
02		21	52.4	+ 31.4
03		22	46.8	+ 24.8
04		27	49.6	+ 22.6
05		28	49.1	+ 21.2

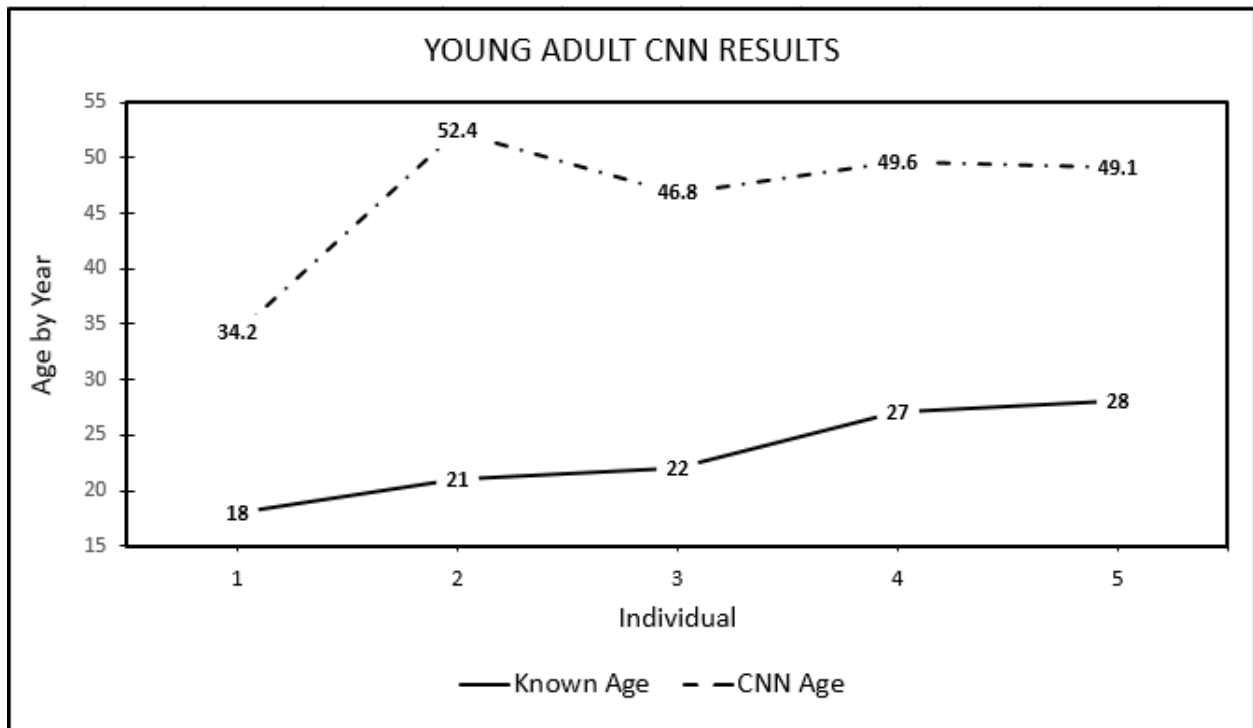


Figure 3.1 Young Adult CNN Results



Table 3.2 Young Adult Observer Scores

Young Adult Observer Scores			
Individual	OBS 1	OBS 2	OBS 3
01	1	1	1
02	4	5	4
03	2	2	2
04	4	4	3
05	3	3	3
Note: incorrect phase placement indicated by shaded cell			

### Middle Adults

Individuals 06 - 12 are classified as Middle Adults. Individuals in this age cohort range from 39 - 62 years old at age-at-death. The MA age cohort represented 56.9% of FANSIE's training set. The CNN was able to produce accurate ages for all individuals. FANSIE's overall error rate in the MA cohort was +1.3. Notably, all individuals were classified within 10 years of known age; 5 of 7 individuals were within 5 years of known age (Table 3.3; Figure 3.2).

Table 3.3 Middle Adult CNN Results

Middle Adult			
Age Range		Training Set %	Mean Error
35 - 66		56.9	+ 1.3
Error Rate by Individual			
Individual	Known Age	CNN Age	CNN Error
06	39	37	- 2
07	40	49.8	+ 9.8
08	47	54.3	+ 7.3
09	50	54.4	+ 4.4
10	57	51.8	- 5.2
11	59	58.6	- 0.4
12	62	57.2	- 4.8

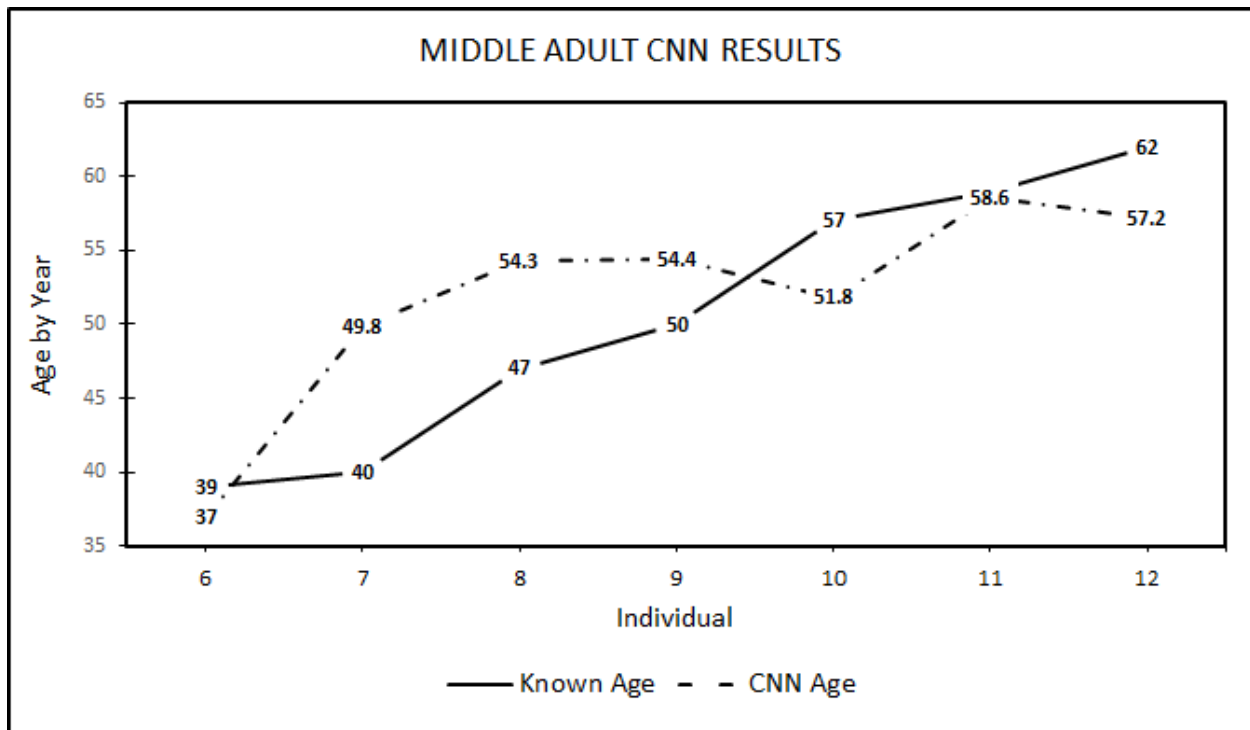


Figure 3.2 Middle Adult CNN Results

Observers were overall accurate in classifying individuals in the MA age cohort (Table 3.4). While there was disagreement among Observers, only reaching unanimous consensus in scoring I08 as phase 4, the known age was within the 95% CI of all assigned phases. In only one instance was an incorrect phase assigned. Individual 11 (known age 59) was underestimated in age by Observer 2, and incorrectly assigned to phase 4 (95% CI 23 - 57).

FANSIE was accurate in all age classifications, and able to produce age determinations within 5 years of known age for 5 of 7 individuals in the MA cohort. Observers made one error in classification. Observers may be regarded as able to produce accurate classifications of age for the MA age cohort. However, due to the broad 95% CIs of phases assigned, and disagreement in scoring, Observer age classification lacks precision.

Table 3.4 Middle Adult Observer Scores

Middle Adult Observer Scores			
Individual	OBS 1	OBS 2	OBS 3
06	4	4	4
07	5	4	4
08	4	4	4
09	5	4	5
10	6	5	5
11	5	4	5
12	6	5	5
Note: incorrect phase placement indicated by shaded cell			

### Old Adults

Individuals 13 - 16 are classified as old adults. The OA age cohort is defined as 67 - 100 years at age-at-death, and Individuals tested from this cohort range in age from 68 - 91 years old. The OA age cohort represents 23.9% of FANSIE's training set. Among older adults, FANSIE underestimated age in all individuals tested, with an overall error rate of -21.2 (Table 3.5, Figure 3.3).

Table 3.5 Old Adult CNN Results

Old Adult			
Age Range		Training Set %	Mean Error
67 - 100		23.9%	- 21.2
Error Rate by Individual			
Individual	Known Age	CNN Age	CNN Error
13	68	53.9	- 14.1
14	68	56.2	- 11.8
15	82	60.7	- 21.3
16	91	53.6	- 37.4

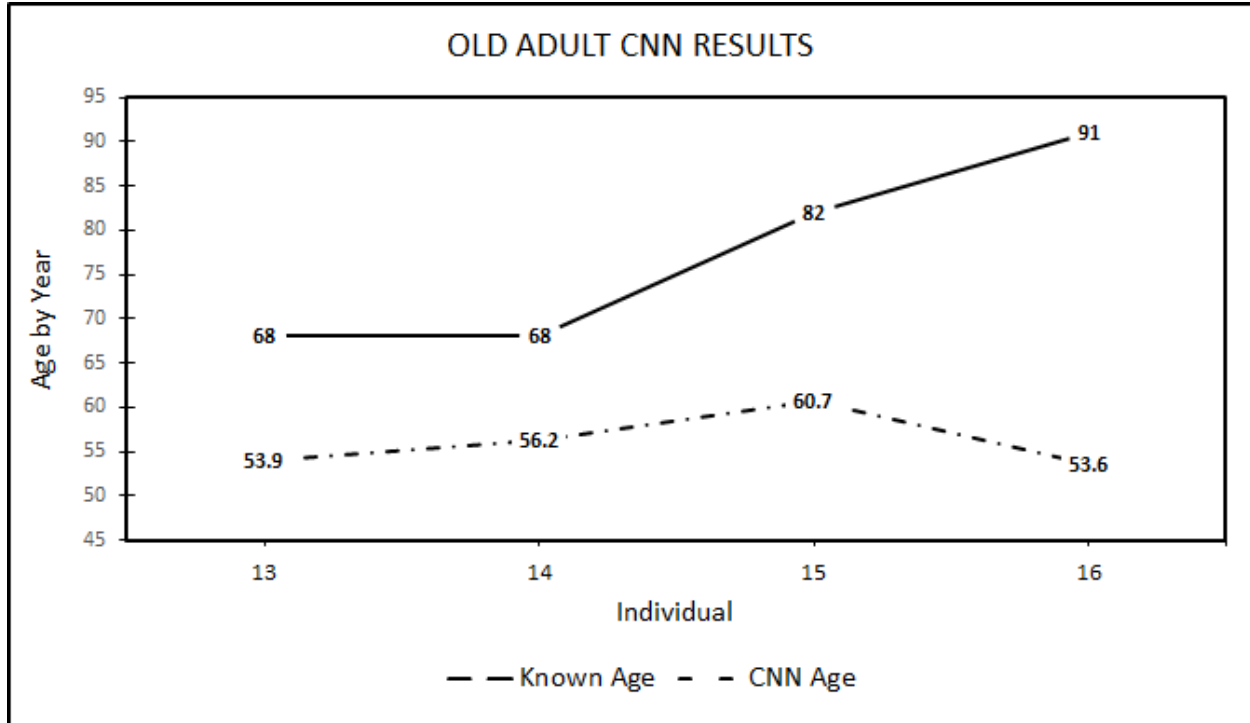


Figure 3.3 Old Adults CNN Results

Observers were only moderately more adept at producing accurate age estimates for Individuals in the OA age cohort (Table 3.6). No individual in this cohort was unanimously classified, or unanimously correctly assigned to a SBP. I13 was underestimated in age by all Observers. I14 was correctly assigned a SBP of 6 in 2 of 3 observations, and I15 was correctly classified in 1 of 3 scorings. I16 was underestimated in age by all Observers. Observers 2 and 3 incorrectly placed I16 in the SBP 5 category. Observer 1 assigned an SBP score of 6. While the Phase 6 category provides the highest age estimation available using this method, the known age of I16 is 91 years, exceeding the limitations of the 95% CI of this phase, and is therefore considered incorrect.

FANSIE was unable to produce any accurate age estimations for individuals in this cohort, and consistently underestimated age. Observers also frequently underestimated the age of Individuals in the OA age cohort, only correctly assigning an accurate SBP for 3 of the 12 observations (25% of all observations). Due to the limitations of the SB method for confidently assigning an age category to advanced geriatrics, the method does not allow for any Observer to

accurately classify I16, as the known age-at-death of 91 years exceeds the upper limit of the 95% CI of 86 years. Observer 1 did assign the highest phase score of 6 to I16, however Observer 2 and 3 underestimated age by assigning a phase score of 5.

Table 3.6 Old Adult Observer Scores

Old Adult Observer Scores			
Individual	OBS 1	OBS 2	OBS 3
13	5	4	5
14	6	6	5
15	6	5	5
16	6	5	5
Note: incorrect phase placement indicated by shaded cell			

## Overall Results

### CNN Results

Average performance of the CNN for all Individuals tested yields an error rate of 14.3 years. However, this average flattens out FANSIE's poor performance in the YA and OA age cohorts, while disguising the CNN's high accuracy age estimation among the middle-aged adults. FANSIE failed to produce an age guess that fell within the prescribed 'young adult' age category for any Individual in the test set, consistently overestimating ages among these Individuals with a mean error rate of + 23.2. This cohort was the least represented in the training set, representing only 18.7% of the training sample. FANSIE was also unable to correctly produce an age guess for any individual in the 'old adult' age category. FANSIE consistently underestimated ages among individual in the OA age cohort, failing to correctly assign an age to any Individual that fell within the prescribed range. FANSIE's mean error for this age cohort was - 19.7, and this cohort was also weakly represented in the training set, accounting for 23.9% of all samples.

FANSIE was able to produce accurate estimations of age for the middle adults. She accurately classified all individuals within the proscribed age range, and was accurate within 10 years for all individuals in the test set. She was accurate within 5 years for 5 of the 7 Individuals tested. The mean error rate for this Individuals in this age cohort was + 1.3. The MA cohort

comprised the bulk of the training set, representing 56.9% of all samples. Table 3.7 summarizes the overall results.

## Observer Results

Observers were overall able to accurately assign a correct SBP score to Individuals in the test set. Observers accurately assigned correct scores for 73% of all observations. Among young adults, observers only misclassified one individual, overestimating age for Individual 02. Observer scores were correct for 80% of all classifications among the young adults. Among middle adults there was only one instance of misclassification; Observer 2 underestimated the age of I11. Observers were less successful at accurately assigning an SBP to older adults. All Individuals were consistently underestimated in age by at least one Observer. Indeed, only 3 of the 12 scores assigned to individuals in this age cohort were accurate. I16, whose known age-at-death is 91 years, exceeds the upper limit of the Suchey-Brooks 95% CI for the highest score (Phase 6, 86 years). For this individual it is not possible to accurately assign a correct score using the Suchey-Brooks method. However, two Observers did place this individual in the lower threshold Phase 5 category, and all Observers underestimated age when classifying with the FSC method, which includes a phase 7 that accounts for geriatric individuals.

There was regular disagreement among Observers. Observers only unanimously assigned the same score to 5 of the 16 Individuals in the test set (I01, 03, 05, 06, and 08), and all such scoring was accurate. Agreement can be regarded as positively correlated with accurate age estimation. Among young adults, disagreement did not result in misclassification. While there was disagreement regarding I02, all Observers were incorrect. Disagreement among Observers regarding I04 did not result in misclassification. Among middle adults only one instance of disagreement resulted in misclassification - Observer 2 placed I04 in SBP 4, underestimating age, while Observer 1 and 3 were correct. Disagreement among observers is positively correlated with advancing age. Half of the test sample, I09 - 116, showed disagreement among Observers, and there was no unanimous scoring among those in the OA age cohort. Disagreement may also be regarded as positively correlated with misclassification for older adults as there was neither consensus nor accuracy in assigning scores among this cohort. Because Observer 1 had no incidences of inaccurate classification of an Individual wherein other Observers were correct,

and was the only correct Observer of I14, Observer 1's scoring is used to represent human evaluation of the material.

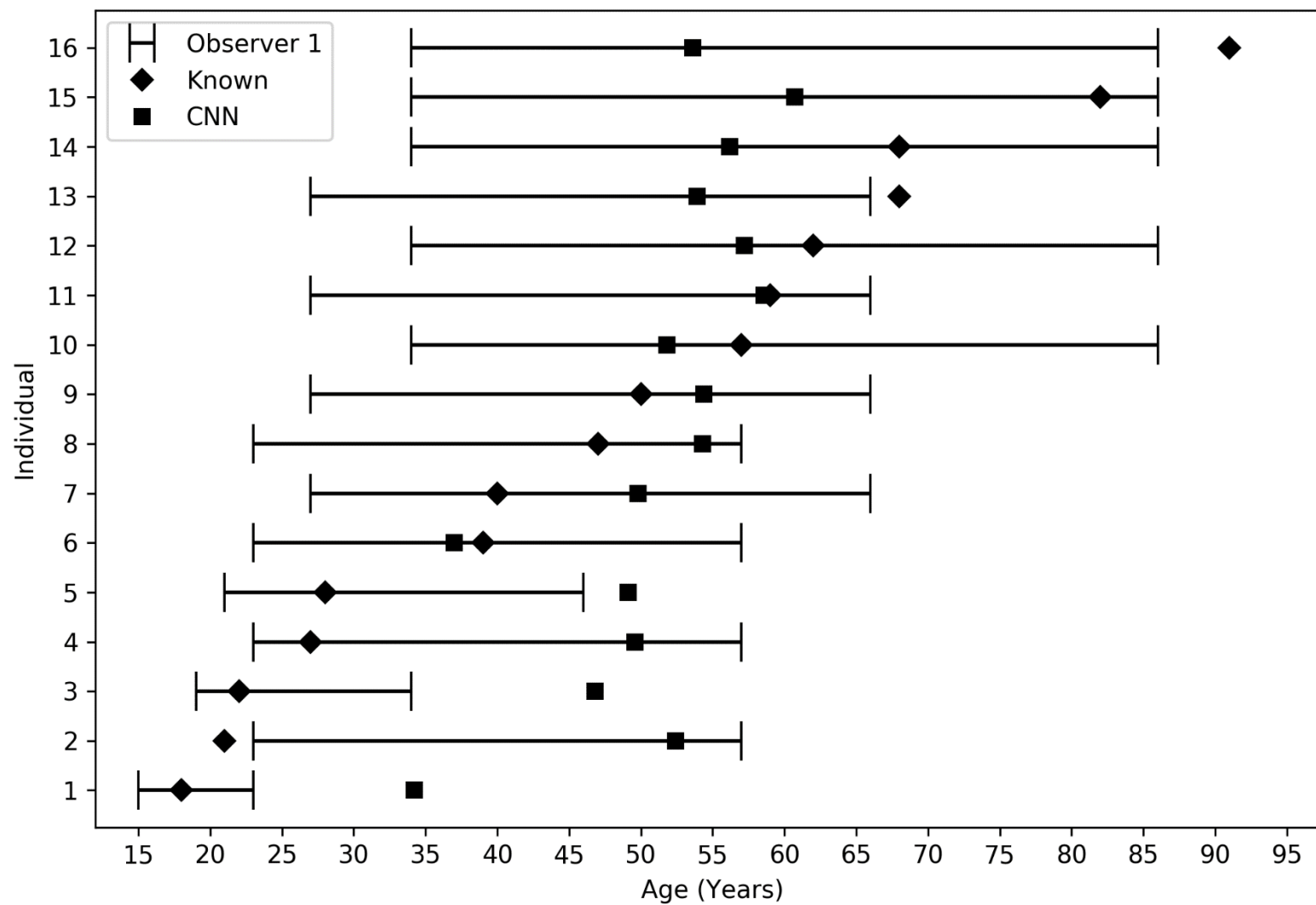
Table 3.7 Overall Results\*: Known Age, CNN, and Observers

Overall Results for CNN and Observers									
CNN Overall Error Rate					Observers Overall Phase Scoring Accuracy				
14.7					73%				
Cohort	Individual	Known Age	CNN Age	Error	Mean Error	OBS 1	OBS 2	OBS 3	% Accurate
Young Adults	01	18	34.2	+ 16.2	+ 23.2	1	1	1	80%
	02	21	52.4	+ 31.4		4	5	4	
	03	22	46.8	+ 24.8		2	2	2	
	04	27	49.6	+ 22.6		4	4	3	
	05	28	49.1	+ 21.2		3	3	3	
Middle Adults	06	39	37	- 2	+ 1.3	4	4	4	95.2%
	07	40	49.8	+ 9.8		5	4	5	
	08	47	54.3	+ 7.3		4	4	4	
	09	50	54.4	+ 4.4		5	4	5	
	10	57	51.8	- 5.2		6	5	5	
	11	59	58.6	- 0.4		5	4	5	
	12	62	57.2	- 4.8		6	5	5	
Old Adults	13	68	53.9	- 14.1	- 21.2	5	4	5	25%
	14	68	56.2	- 11.8		6	6	5	
	15	82	60.7	- 21.3		6	5	5	
	16	91	53.6	- 37.4		6	5	5	

\*Errors in estimates highlighted



Figure 3.4 Overall Results: Known Age, CNN, and Observer 1



## CHAPTER IV

### DISCUSSION

#### Integrating and Interpreting Machine Learning Systems

##### The Black Box Problem

One of the great challenges associated with integrating machine learning approaches to sensitive contexts, such as medicine or the legal process, concerns the issue of what is termed in the parlance of computer sciences as 'interpretability'. While interpretability has been defined in different ways throughout the scholastic discourse, a broadly applicable definition devised herein by the author is simply the capacity of a human to make meaning of a machine learning system's output. This issue of interpretability may be particularly problematic to resolve in approaches that utilize complex machine learning techniques, such as the CNN employed in this research, which operate non-linearly and are laden with hidden dimensions (the deep layers of deep learning). FANSIE has 20 hidden layers, each one of which responds to input data in discrete and divergent fashions, ultimately yielding a complicated miasma. The byzantine guts of complex machine learning systems are difficult at best for humans to understand. They are entirely impossible to understand in the manner of the machine.

CNNs, and other machine learning strategies, are famously mysterious in their execution. While the input will be known and understood by humans (e.g. the test set, or the training set), and likewise the output can also be understood (i.e. the age determination produced), the steps in between are termed opaque, and may be potentially wholly indecipherable. Even the programmer who has created the system will likely be largely ignorant of the processes occurring in the layers of the network that inform its analysis of data and dictate its output. The opacity that characterizes the inner workings of machine learning systems is known as the black box. Input is

received, the network analyzes the data in some obscure fashion, 'the black box', and then a human decipherable output is produced.

As it concerns the research presented in this thesis, the input is readily understood. The volume of scholarly research concerning the pubic symphysis is legion, and the training set used is rich with attendant metadata. The precise technique and nature of the data set's digitization is well documented. Likewise, the output is an unambiguous and plainly stated real number which is FANSIE's best guess at classifying the input data according to the method proscribed by the human programmers, a linear regression function. FANSIE's architecture, the code which serves as the digital machine, is also readily understood having been constructed according to the specifications of its creators. However, the chain of events that follows from the introduction of the thoroughly understood input and through the layers of FANSIE's well-defined architecture, to the resulting plainly stated output is a rather murky affair.

The nature of the black box also complicates the task of human analysis of results. It is a straightforward matter to assess the *what* of FANSIE's results - that is to say, how accurate they are. But, it is a rather more complicated task to tease out the *why*, those aspects of morphology that are mathematically construed as diagnostically relevant and thus inform output classifications. First, a brief examination of the integration of machine learning into forensics from both socio-institutional and legal perspectives is presented. Next, the black box problem is evaluated as it concerns interpretability by exploring the practical, critical, and philosophical foundations of meaning-making, while offering pragmatic strategies for interpretation.

## Challenges to Integration

### Practitioner Bias

The enigmatic nature of the black box creates some potential obstacles to incorporating the use of FANSIE, or any other machine learning system, to bioarchaeological research and especially applied human osteology in forensic scenarios. Two important impediments to the successful integration of machine learning approaches in anthropology are considered here. First, integration is challenged at the level of the practitioner and their framing institutions, which may be understood generally as the practitioner's foundational traditions, such as a particular

academic milieu, rather than as a formally defined space or entity. Lipton (2017) suggests that users, though referring specifically to medical practitioners, may place more trust in systems that are prone to the same sorts of errors that humans make when analyzing data, but are very distrustful of systems that make mistakes where humans would not.

The important implication for integration into practice is that poor performance in any dimension of analysis creates distrust in systems as a whole, regardless of the quality of other analyses. Lipton additionally states that there is a mistrust of systems that perform better than humans in complex tasks. So presumably a system that is trusted, and may therefore actually be used, is a system that performs only as well as a human! Meanwhile, systems that outperform humans can be regarded with suspicion and are unlikely to be incorporated into practice, despite the benefits they may confer. Lipton magnanimously dubs this technophobia as "institutional biases against new methods" (pg. 7, 2016). These biases are almost surely characteristic of training that emphasizes cautious approaches, which certainly includes medicine but probably also extends to forensic anthropology.

Providing users with some means of understanding the network's mechanism of operation may promote greater trust in the results it generates. Overcoming this challenge to machine learning integration will be daunting, however, as creating systems that are more transparent in their operations often comes at the cost of predictive power. Complex models produce better results, but are also more opaque as a consequence of increasing system intricacy associated with large data sets. While sacrificing performance for integration may diminish the potency that makes approaches like neural networks a powerful tool, a system that isn't being applied to the tasks it was developed for is hardly useful. Mitigating the tensions between these contrary objectives may well need be addressed on a case-by-case basis as there are neither established nor tentative disciplinary standards for either acceptable degrees of accuracy or acceptable parameters of system transparency.

### Legal Admissibility Standards

The second impediment to machine learning integration concerns the potential legal complications of using this technology in forensic contexts. Machine learning is novel enough in the forensic sciences that research is still in early stages, and as of the time of this writing not yet

being actively applied to any medicolegal case work. However, given the rapid proliferation of machine learning approaches to a variety of other industries, it seems inevitable that they will be eventually utilized in some forensic fashion. Therefore, it is prudent to provide some moderate considerations of how machine learning may be approached as a matter of jurisprudence.

The Daubert standards, established via the 1993 Supreme Court ruling in the landmark *Daubert v. Merrell Dow Pharmaceuticals, Inc.* establish the criteria under which expert testimony may be deemed admissible in court. The Daubert case reified the congressionally enacted Federal Rules of Evidence, while dismissing the previous standard of scientific admissibility, the so-called 'Frye test', which established as admissible only that which had met general acceptance within the relevant community of experts (*Frye v. United States*, 1923). Daubert standards are often understood to serve as a litmus against which the legitimacy of some aspect of science may be gauged, and in a strictly legal sense that is correct. The language of the ruling, however, is remarkably unclear as to what actually constitutes either an expert opinion or valid science. Instead, the Court offers suggestions for establishing the veracity of scientific notions which the presiding judge may or may not choose to follow. In their dissent, Chief Justice Rehnquist and Justice Stevens describe the suggestions offered by the Court as "vague and abstract" (*Daubert v. Merrell Dow Pharmaceuticals, Inc.*, 1993), among other similar criticisms.

By design, the Daubert ruling affords great flexibility and a broadly liberal approach to the admissibility of expert testimony and matters of science. Indeed, in their brief, the concurring Justices explicitly state that while peer review and publication may serve as an appropriate metric for establishing scientific validity, there is no inherent mandate that scientific methods or knowledge undergo a peer review process or be published to be deemed admissible. The only truly essential criterion is that testimony is germane to the particularities of the case. Rather than explicitly define a set of parameters under which science and scientists may be evaluated, the ruling instead grants the presiding judge the authority to assess the legitimacy of expert testimony, methods, and technique and to further allow or deny the same as they deem appropriate.

Here, then, a nascent complication of machine learning applied to forensic purposes ensues. Medicine and law, which tend to be relatively cautious adopters of technology, are

married in the medicolegal framework within which forensic anthropology operates. Naturally, the computer sciences and associated disciplines are much quicker to adopt emerging technologies. There is potential for discordance among experts wherein a computer scientist would very likely argue that machine learning generated data is wholly legitimate and appropriate for admissibility, while a traditionally trained anthropologist, suspicious of results seemingly too accurate, precise, or random to be correct, argues the contrary.

Ideally, as well as most likely, all parties involved in a forensic investigation will be in accord before cases are tried and the adversarial role of refuting an expert's testimony will be the purview of attorneys, rather than scientists, however there are numerous examples of conflicting expert opinions in medicolegal investigations. The Daubert case itself represents just such a scenario. Should a disagreement occur in a case involving the use of machine learning, it is uncertain which camp the relevant expert should hail from, or whose opinion is best suited to meet the legal needs of the case - the technological, or the anthropological? Machine learning advocates, attempting to convince either a reticent judge *or* a reticent discipline, are faced with the mutual dilemma of how best to promote understanding and acceptance of technological processes, that through virtue of their function, may only be considered in the gestalt and defy reducibility.

### Strategy for Interpretation

In order to establish an appropriate strategy for interpretation of neural networks in general, and more especially to FANSIE's results, it must first be determined what it is *exactly* that requires interpretation. The concept of 'algorithmic opacity' has been subject to the same sort of academic scrutiny as algorithmic interpretability, and is likewise often employed nonspecifically. In this manuscript, opacity has been used as a descriptive proxy for the black box. Qualifying what is meant by the black box beyond merely the 'opaque processes' of some machine learning system is therefore mandated. Using these terms synonymously, while intuitively convenient, is ultimately tautological and therefore not particularly meaningful. A black box is an opaque algorithmic process, and an opaque algorithm process is a black box. Burrell (2015) clarifies these ambiguities by unpacking the meaning of opacity in different

machine learning contexts. She identifies two forms of algorithmic opacity that are relevant to interpreting FANSIE.

The first aspect of opacity concerns the gap in technical literacy between the lay and specialist communities. Writing code and computer programming are highly technical skills that, as in any specialized trade, require substantial education and training to perform. While there is an increasingly pervasive ubiquity to algorithmic approaches to all manner of common tasks, the spectacular pace of technology has left public education in the dust. The resulting lacuna in popular knowledge has produced a certain cultural zeitgeist that regards many aspects of complex technology as inherently inscrutable. Burrell posits that this is simply a consequence of educational deficiencies that are entirely resolvable, and others propose changes to traditional educational paradigms in favor of approaches that promote technological literacy to address these deficiencies (Lee, 2011; Wing, 2006).

A more immediate, though imperfect, resolution to this type of opacity is to allow open access to source code, and therefore open access to public examination. Of course, the ability to view the code does not provide the technical expertise necessary to make sense of it. However, it does offer what is the most unambiguous and transparent look at FANSIE, if not the most understandable. As a bonus, the ethical considerations inherent in the use of secret strategies to produce sensitive and personalized information about an individual are reduced, while also presenting the opportunity for any interested party to examine, modify, or augment the code in productive ways. In this sense, FANSIE may be regarded as interpretable. Interpretability, as defined here, states only that a human has the *capacity* to create meaning, but not necessarily that they do. All of the components that comprise FANSIE may be freely examined by any who so desire. Indeed, by unanimous agreement among myself and my project collaborators, it was decided very early in the project planning that FANSIE's source code would be freely shared online, to best serve the interests of scientific progress, ethical dictums for transparency, and the forensic community.

Of course, for most people a virtual tome of code will be neither interesting in and of itself, nor especially meaningful. Regardless, it does serve to highlight the substantial difference between that which is unknown only because of differential education and training, and that which is innately unknowable as a product of fundamental incompatibilities between human and

machine modes of cognition. This is the crux of the second, and more essential form of opacity defined by Burrell, who summarizes thus, "Machine optimizations based on training data do not naturally accord with human semantic explanations" (pg. 5, 2015).

A comprehensive explanation of a CNN's mechanism of operation may be found in Chapter I: Introduction, Introduction to Machine Learning, but a concise summary will now be offered to better contextualize the discussion that follows. Succinctly, a CNN may be understood as a system of data classification wherein the input data is transformed into a matrix of values which are interpreted differentially, relative to the weight of those values, such that the differences among them progressively and accumulatively inform either the further transformation of that data, or its eventual output into some defined category that is deemed as best fit according to the CNN's previous assessment of category exemplars. The manner in which the CNN chooses to levy more or less weight, which is to say how the system decides which areas are important in classification, and which areas are not, is entirely of its own devising. Further, while it is at least technically feasible to follow data through all the layers of the CNN, from input to output, and document the evolving nature of the weighted matrices, such efforts do not bear fruit that is meaningful to humans.

This presents the dilemma of how to produce an explanation of CNN output that is both satisfying to a human, and predicated upon a legitimate evaluation of the network's activities. To offer interpretations of FANSIE's output, a case based approach of post-hoc examination was developed. It has been noted that users may regard their understanding of a CNN's output as satisfying if given explanations that are based on case similarity. Caruana *et al.* state "With case-based methods, practitioners often are satisfied with explanations that consist of cases that the model judges to be most similar to the test case." (pg. 212, 1999). This is particularly interesting because this kind of 'explanation' does not truthfully explain anything at all, but is still meaningful to the user.

The definition of machine learning interpretability proposed herein was developed with careful semantics for this reason. That is, that a human may 'make meaning' out of what a machine outputs. The system is therefore interpretable, while simultaneously being neither explained nor understood. It seems likely that in such circumstances the user is generating post-hoc explanations based on the model's output. Fortunately, for image based machine learning



systems, such as CNNs, it has been proposed that post-hoc interpretation is, in fact, an appropriate technique. Per Lipton, "neural networks tend to operate on raw or lightly processed features. So if nothing else, the features are intuitively meaningful, and post-hoc reasoning is sensible." (pg. 7, 2017). An approach of case based interpretation based on assessment of morphologically salient features, and corroborated with test set examples, was subsequently developed to evaluate FANSIE's output. Observer scoring was applied to further contextualize the results.

## Individual Case Studies

### General Context for Case Studies

FANSIE is poor at estimating ages among younger adults and older adults. Young adults are consistently overestimated in age. FANSIE failed to accurately classify any individual in this cohort, and young adults are under-represented in the training set. With unique exception, Observers are able to accurately categorize young adults into an appropriate SBP. The exceptional individual among the young adults, I02, was also notably over-aged by FANSIE. Observers are only marginally better at classifying older adults than the CNN. Older adults are consistently under-aged by FANSIE, and no individuals in the Older Adult cohort were classified correctly. Observers also tended to underestimate ages among the elderly, and there was less agreement in classification among them regarding this cohort. For advanced geriatrics, such as I16, the limitations of the method do not allow for accurate classification as the upper threshold for 95% confidence terminates at 86 years.

The CNN is able to produce very accurate ages for individuals in the middle adult age cohort, which comprised over half of the total training set. Mean error in this cohort was +1.3 years, and FANSIE was accurate within 10 years for all individuals, within 5 years for 5 individuals. Observers frequently disagreed about score assignments among the middle aged, but only one score was inaccurate. Both the CNN and Observers are competent at accurately providing age estimations for middle adult Individuals. Due to the wide ranges and broad overlap of phases assigned in this cohort, however, FANSIE is considered better able to produce accurate ages among middle aged adults than Observers.

## Individual 01

Individual 01 has a known age-at-death of 18 years. This individual was unanimously, and correctly, assigned to the SBP 1 category by all observers. The CNN overestimated the age of I01 by 16.2 years, out of bounds of the 95% CI of the Phase 1 category, and beyond the threshold of the defined young adult age range (Figure 4.1).

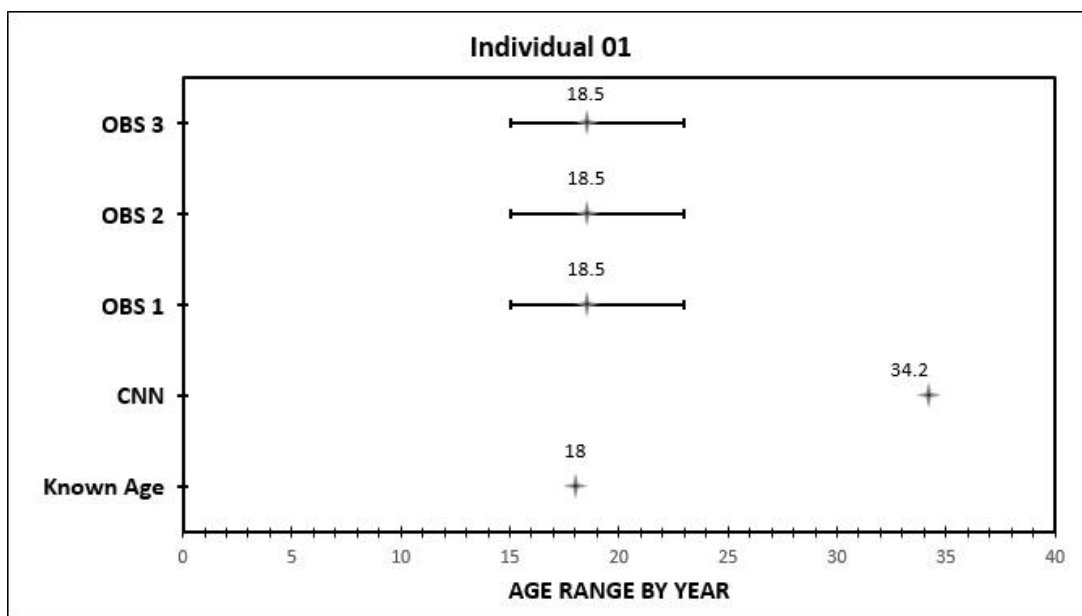


Figure 4.1 Individual 01: Known Age, CNN Age, and Observer Scores

Figure 4.2 shows a photograph of the real bone that Observers were tasked to classify on site, and an image of the 3D scan that was provided for FANSIE's review. The face of this Individual exhibits cortical bone damage to most of the upper extremity and ventral aspect. The Observers are able to extrapolate from the undamaged portion of the face along the dorsal aspect of the lower extremity that the deep furrows characteristic of youthfulness are representative of the whole feature and thus correctly determine that I01 is from a very young adult.

Due to the cortical bone damage, laser scanning produced a poor representation of I01. The bony crypts of the exposed trabecular tissue tend to scatter laser light within their cavities rather than reflecting back to the device. This is interpreted as absent data. The voxelization process smooths out much of the absent data, as the dimensions of each voxel exceed the dimensions of the gaps. However, the region is flat and generally featureless relative to the voxel

dense dorsal aspect (see Figure 4.3). FANSIE may have interpreted the feature dense dorsal region as remnant billowing, and the flattened upper extremity as smooth and fine grained worn bone. This is a morphological patterning sometimes seen among late young adults and throughout the middle ages, as demonstrated in Figure 4.4 depicting individuals from the training set.

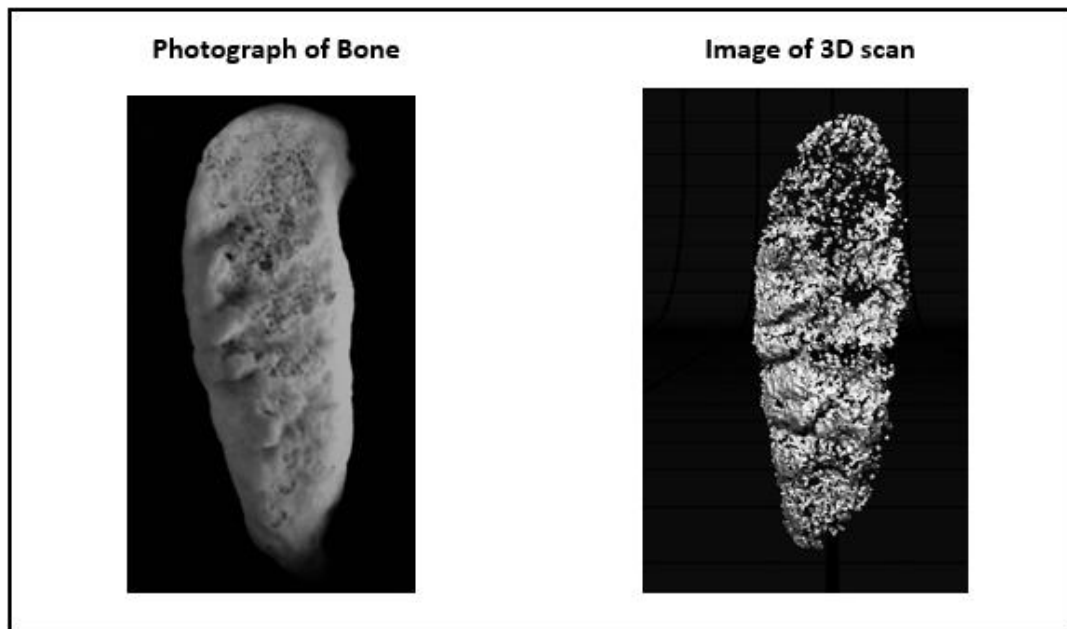


Figure 4.2 Individual 01: Images of Pubic Symphyseal Face

However, it is also possible that the CNN did recognize the deep billowing, and associated that feature with young age, but interpreted the flattened area in the upper extremity in a contrary way, and split the difference. In fact, this youngest Individual in the test set is the only Individual that the CNN generated an age younger than the 40s for (excepting I06, discussed in 'Case Studies' later). So, perhaps there was some recognition that the deep billowing feature is indicative of youth.

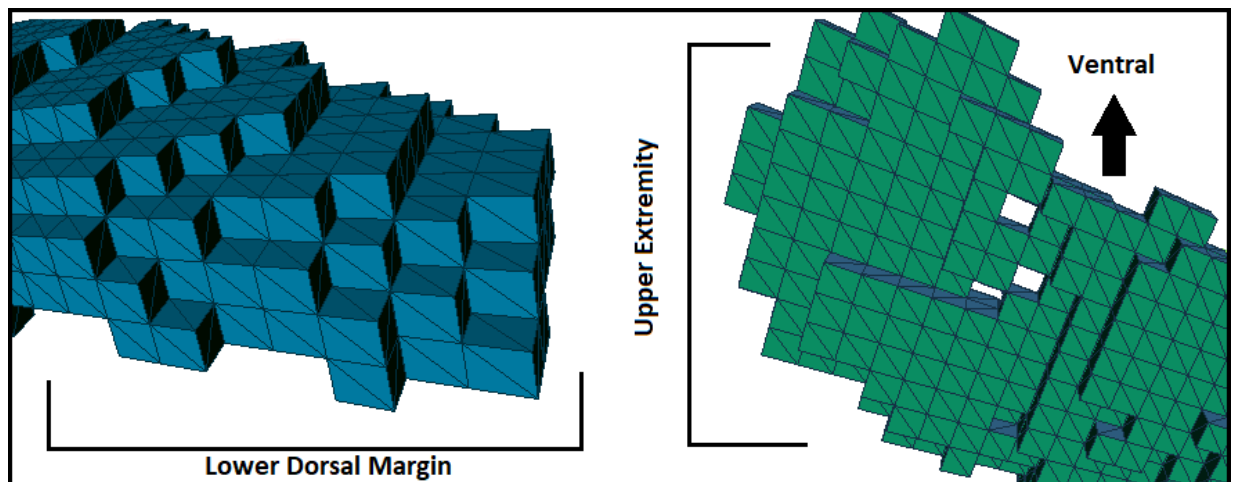


Figure 4.3 Individual 01: Voxelization of 3D Scan

However, it is also possible that the CNN did recognize the deep billowing, and associated that feature with young age, but interpreted the flattened area in the upper extremity in a contrary way, and split the difference. In fact, this youngest Individual in the test set is the only Individual that the CNN generated an age younger than the 40s for (excepting I06, discussed in 'Case Studies' later). So, perhaps there was some recognition that the deep billowing feature is indicative of youth.



Figure 4.4 Representative Images from Training Set Displaying Remnant Billowing

Due to the bone damage displayed on the randomly selected Individual, and the probable resultant misinterpretation, it cannot be determined if the CNN has learned to recognize the distinct morphology of very young adults. Tasking FANSIE to produce an age on an exemplary sample, with an undamaged face and typical morphology, would be the most appropriate strategy to evaluate if the network has learned to make associations between furrowing and youth.

This case demonstrates that skilled human interpretation will continue to be an essential component of forensic skeletal analysis, regardless of the future performance of machine learning systems. The characteristic and distinct PS morphology of very young adults, such as the teenaged I01, was anticipated to be an easy interpretation for the CNN. It is reasonable to assume that the damage to the PS face of this Individual influenced FANSIE's decision. In real world conditions preservation is likely to be non-ideal. Therefore, human interpretation, and the extrapolations that come intuitively with the human manner of pattern recognition, will likely always be necessary to provide essential context, even for high performing CNNs.

## Individual 02

Individual 02 has a known age-at-death of 21 years old. FANSIE produced an age of 52.4, while Observers scored I02 as either SBP 4 or 5. Despite the high age provided by FANSIE, the value still falls within the 95% CI of both assigned phases which have broad overlap and span over three decades (see Figure 4.5). In this case, both FANSIE and the Observers overestimated the age of the Individual.

The facial morphology of I02 displays largely eroded remnant furrowing along the dorsal aspect of the lower extremity, complete delimitation of both extremities, mild depression of the face, and an emerging rim along the ventral margin of the lower extremity (Fig. 4.6). All these features are characteristic of more advanced bony degradation and remodeling than would typically be expected of a 21 year old individual. Observer scoring, though in disagreement regarding phase, is overall consistent with the morphological presentation of this feature.

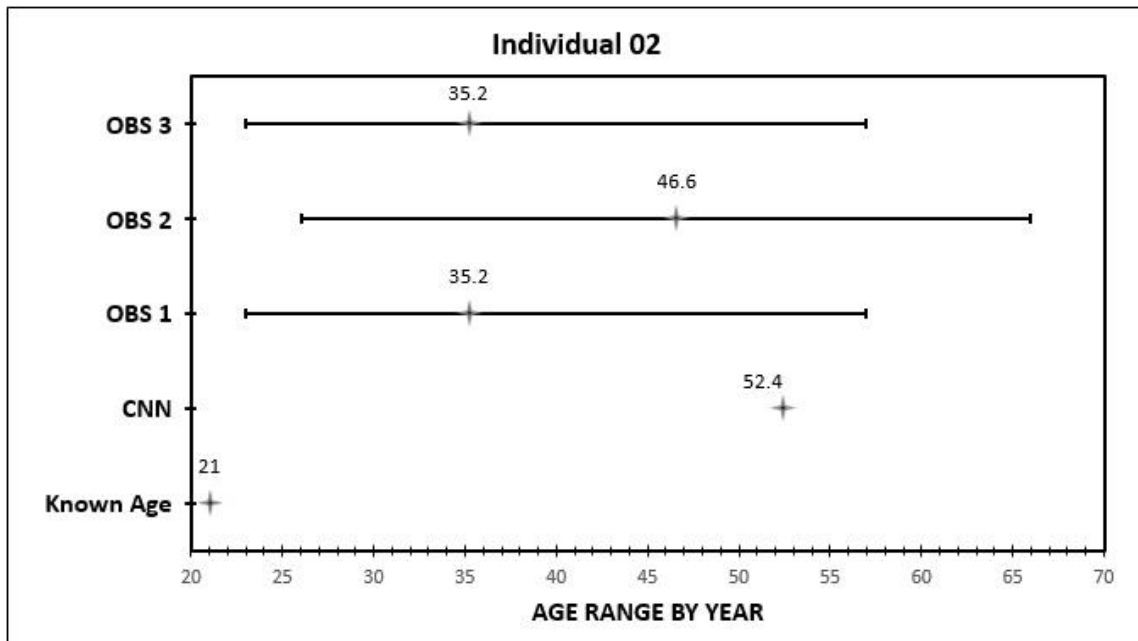


Figure 4.5 Individual 02: Known Age, CNN Age, and Observer Scores

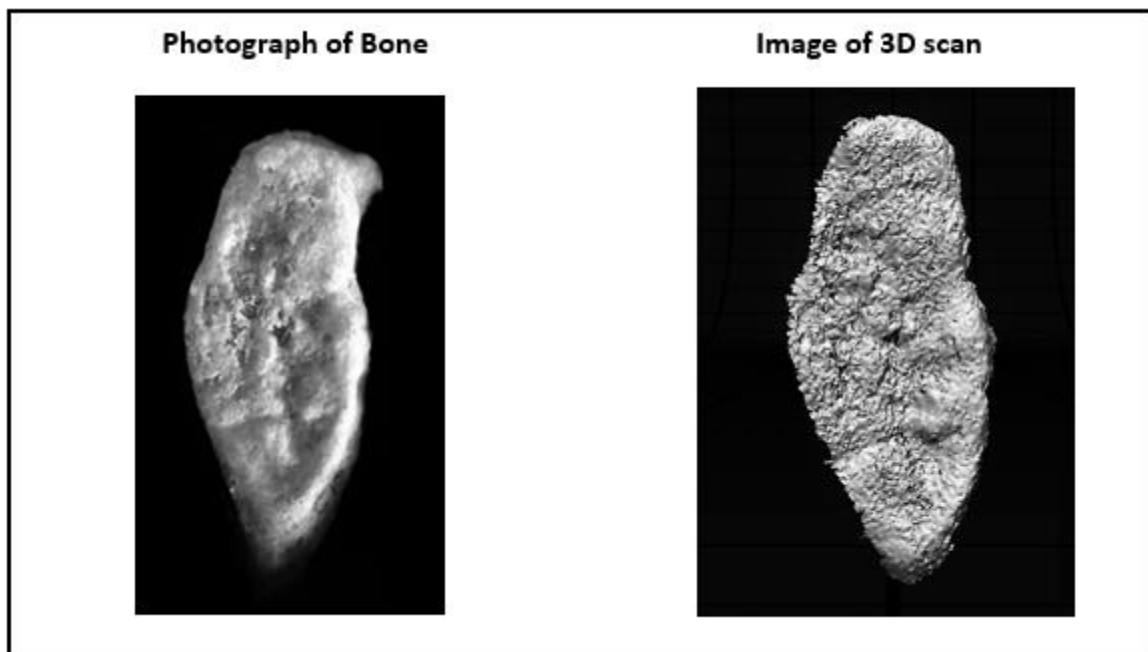


Figure 4.6 Individual 02: Images of Pubic Symphyseal Face

FANSIE produced the highest age determination among the young adults for I02, exceeding even the age determinations for the older individuals in the YA cohort. FANSIE's age

for I02 also has the highest error rate among all Individuals tested, at + 31.4 years, excepting I16, discussed in 'Case Studies' later. The higher age estimate for this Individual, particularly when contextualized with the overestimated SBP scores provided by the Observers, is an important litmus for CNN training and learning potential. It implies that FANSIE has learned to associate some of the features seen in the bone with older ages. Like human observers, FANSIE characterizes this bone as displaying features typical of the middle ages, rather than the true young age of the Individual. It demonstrates that the CNN is making some moderate associations between notable indicators of advancing PS bone degradation and advancing age. This case provides reassurance that the age determinations being produced are predicated upon some recognition of certain features being correlated with particular ages. It also highlights an important limitation of applying machine learning technology to osteology - the essential nature of morphological diversity may preclude high accuracy as the best predictive strategies will still be challenged by inevitable outliers.

#### Individual 06

Individual 06 has a known age-at-death of 39 years. FANSIE determined the age of I06 to be 37 years. All Observers were in agreement and accurately classified this Individual into the SBP 4 category (Figure 4.7). I06 is an intriguing case study for a number reasons. First, this Individual marks the beginning of FANSIE's high accuracy period. I06 is the first among the defined middle adult cohort, where the CNN results are substantially closer to known age than among younger and older adults. In fact, the age estimate for I06 breaks the trend established with I03 - I05, who were all rather erroneously given 46+ ages. Following I06, FANSIE resumes assigning 46+ ages. A tendency to produce middle age guesses for all individuals in the test set, from young to old, will inherently be most accurate among that cohort, and an interpretation of that behavior will follow in the 'Broad Considerations' section of the 'Discussion' chapter.

However, that I06 and I01 are the only Individuals to be assigned ages below 46 years, implies there must be some particularly resonate feature they share, which others in the test group lack, that is associated with the relatively younger age guesses. The break from pattern, as in the case of I02, indicates that associations between morphology and age are being made by the CNN, rather than random assignments that give an impression of accuracy due to coincidence.

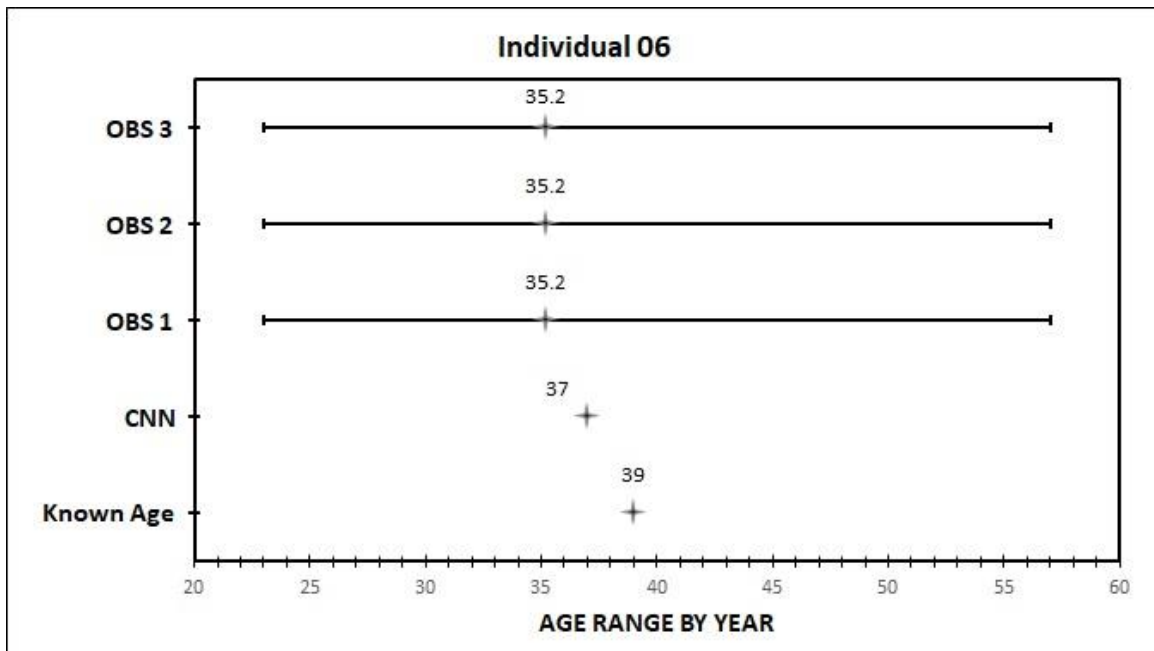


Figure 4.7 Individual 06: Known Age, CNN Age, Observer Scores

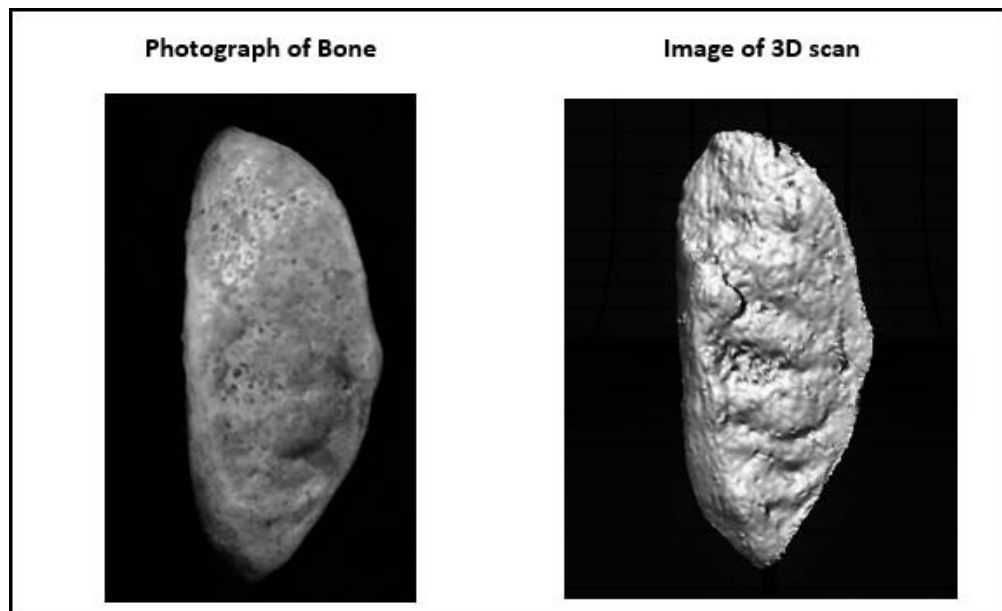


Figure 4.8 Individual 06: Images of the Pubic Symphyseal Face

The features depicted in the photo and the scan bear little resemblance to I01 in terms of human reckoning, though the remnants of billowing along the lower dorsal surface are suggestive. It's important to recall that only speculation about the decision-making process of the



CNN is truly possible, however. So, the strategy of post-hoc feature based analysis can only guide interpretation so far before inevitably butting up against results that challenge assumptions of CNN cognition. Ultimately, FANSIE may or may not learn associations that seem intuitively obvious to humans, while at the same time correlating features than are dismissed as insignificant or irrelevant. The voxelized scan of I06 does show high activity in the lower extremity, as does the lower extremity of I01, which is representative of the remnant furrows of I06 and robust, but damaged, furrowing of I01 (Figure 4.9).

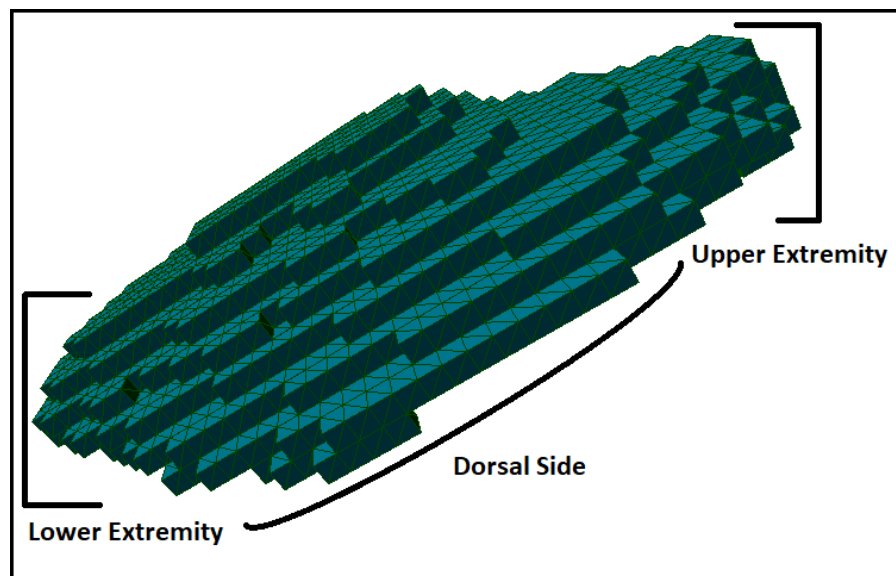


Figure 4.9 Individual 06: Voxelization of 3D Scan

The most intriguing aspect of Individual 06, however, especially considering the high accuracy with which this Individual was categorized by the CNN, is the matter of I06's contrary sidedness. Unlike all other Individuals in the test set, I06's scan is of the right, rather than left, pubic symphysis. This is the product of mislabeling during the initial data collection and coding process, rather than by design. Careful documentation of all right sided scans to avoid just this scenario was attempted during scan labeling, but it is inevitable that some mistakes will occur when processing many hundreds of files. In fact, one of the images used as demonstration in the I01 case study (Figure 4.4) also shows a right side pubic symphysis that was missed during coding and put into the training set.

The total number of right side scans taken is few compared to the left, as this was only done when the left aspect was extensively damaged or otherwise unavailable for scanning (n=12). There are likely other mislabeled right aspect scans hiding in the training set, though probably very few. How much influence these scans may have had on FANSIE's training is uncertain. In this instance it was a happy accident, as it affords an opportunity to examine how the CNN might respond to a scan with mirrored anatomy. The high accuracy of FANSIE's age output, especially when coupled with the somewhat anomalous attribution of a sub 40s classification, implies that whatever features are being associated with the younger, and very accurate, age assignment are not dictated by sidedness.

Capabilities of this nature are a fairly reasonable expectation as one of the key advantages of a CNN's image recognition strategy is the identification of meaningful, focal structures within the whole. Recognizing key structures is what gives CNN's success when identifying and sorting images which have no exemplar to serve as a template and numerous and varied presentations, such as handwriting. FANSIE's success with guessing the age of I06 suggests that some of the features being accurately recognized are generalizable to both left and right sides. Potentially, there is some aspect of the face, divorced from the specific anatomic markers of sidedness, that is particularly meaningful to the CNN. If so, isolating that feature may provide useful context for osteologists conducting age estimation by ways of gross visual inspection of the PS. Voxel dimensions serve as a reasonable proxy of digital resolution. The 38x38x38 boxing that was used in this project renders too large to represent minute details. Therefore, anything that FANSIE sees, a human can also see. If FANSIE is able to make some especially striking association between morphology and age, as seems to be the case here, it would be useful to identify that feature so that it can be incorporated into age estimation techniques.

#### Individual 16

Individual 16 has a known age-at-death of 91 years. This is the oldest Individual in the test set. The CNN produced an age determination of 53.6 years for this Individual. I16 is the least accurately aged Individual in the test set, with an error of - 37.4 years. Two Observers underestimated the age of I16 by assigning an SBP score of 5, which has a 95% CI upper

threshold terminus of 66 years. One Observer assigned this Individual to phase 6, the highest possible with the SB method, which extends the 95% CI range to 86 years (Figure 4.10).

Because the phase 6 category is the highest possible to assign using the Suchey-Brooks method, based on this approach alone it would be impossible to determine if Observer 1 truly underestimated the age of I16, or if it is an artifact of the innate limitations of the method. Fortunately, Observer data was collected which resolves this dilemma. During scoring, Observers were asked to place the test samples into the appropriate phases using two methods.

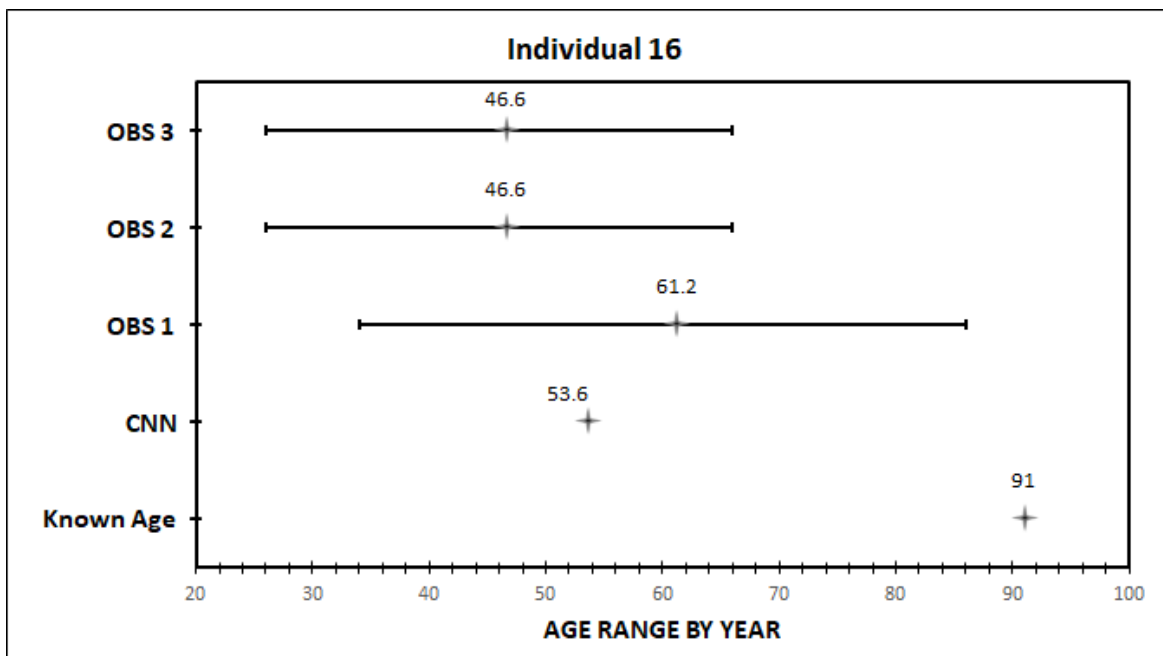


Figure 4.10 Individual 16: Known Age, CNN Age, Observer Scores

First, Observers scored using the Suchey-Brooks method that is the disciplinary standard, and has served as the foundation for evaluation of human performance in this research. Observers were also asked to score using the FSC method which was developed from the same population evaluated here. Observers had training and experience with both methods before recruitment, and currently use the FSC method as their standard age estimation technique.

Integrating the FSC method scores was ultimately deemed beyond scope for this project, as the primary focus of the research is the performance and utility of the CNN, but here it is referenced to allow for better nuance in the discussion of I16. In this case, all Observers

unanimously placed I16 into the FSC phase 6 category. This phase has an age range of 51 - 83, and so assignment in this phase still underestimates the age of I16 (Figure 4.11). This is meaningful because age estimation with this method provides Observer consensus that still unanimously underestimates the age of I16, as with the SB method. Further, the FSC method includes a phase 7 category that encompasses ages from 58 - 97 years. Therefore, the underestimation of age by all Observers can be understood to represent authentic interpretation of PS morphology, rather than merely being secondary to the limitations of the method.

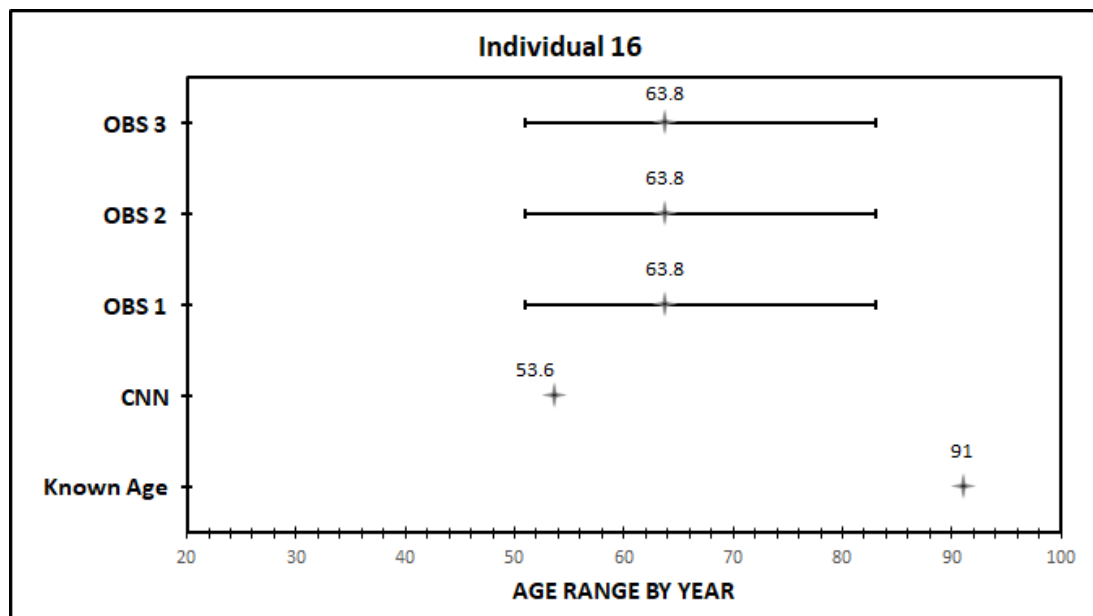


Figure 4.11 Individual 16: FSC Method Age Estimation Results

The CNN produced a rather inaccurate age determination. However, as in other case studies these results actually reify CNN competency despite the high error margin from the known age-at-death. Much like I02, in this circumstance we have an Individual whose PS morphology is notably divergent from the expected norms of their age. I02 displayed surprisingly degenerated surface morphology relative to their rather young age. Here, I16 shows remarkably resilient bone structure relative to their rather advanced age (Figure 4.12).

The facial morphology of I16 is consistent with an older, but not advanced geriatric, individual. There is neither development of lipping nor breakdown of the rim, the superior ventral aspect is still well-defined, the shape is regular, and the face displays no pitting or

macroporosities. Particularly when compared to other advanced geriatrics (as seen in Figure 4.13, showing individuals from the training set), I16 has a relatively youthful appearance.

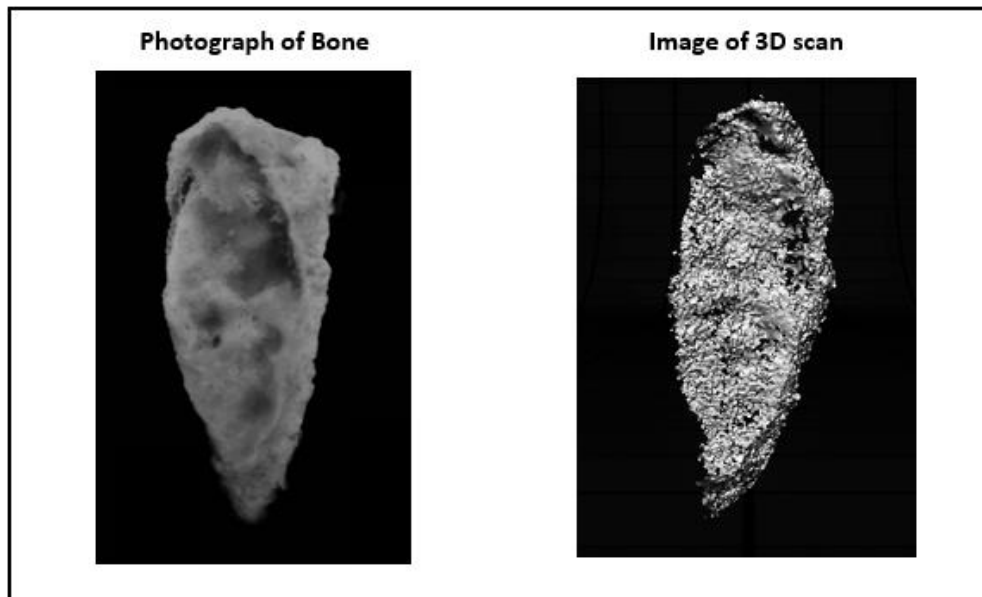


Figure 4.12 Individual 16: Images of the Pubic Symphyseal Face

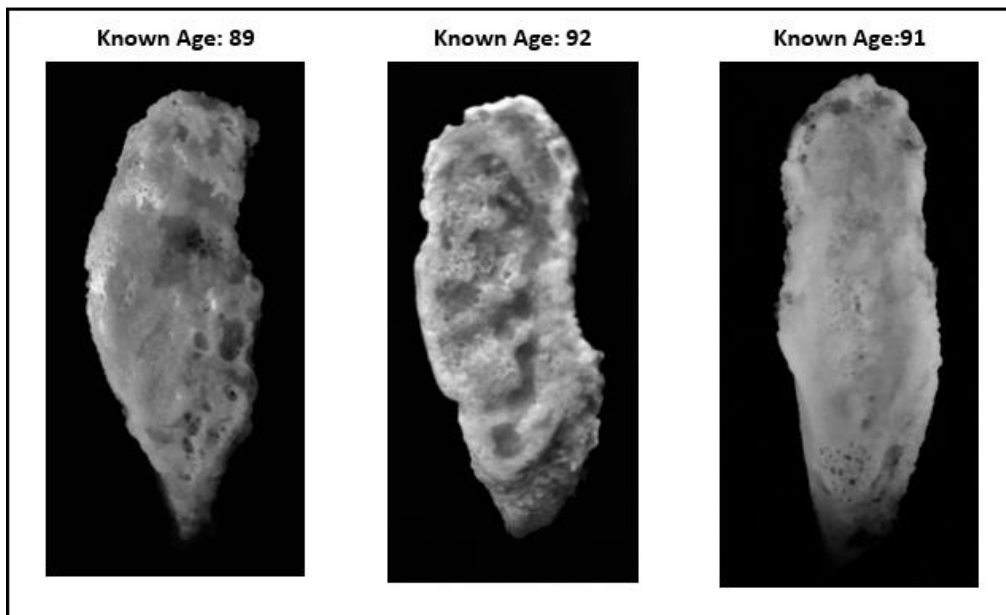


Figure 4.13 Representative Images from the Training Set, Geriatrics

Figure 4.14 shows the facial morphology of I15, with a known age-at-death of 82 years, and the oldest CNN estimate at 60.7 years.

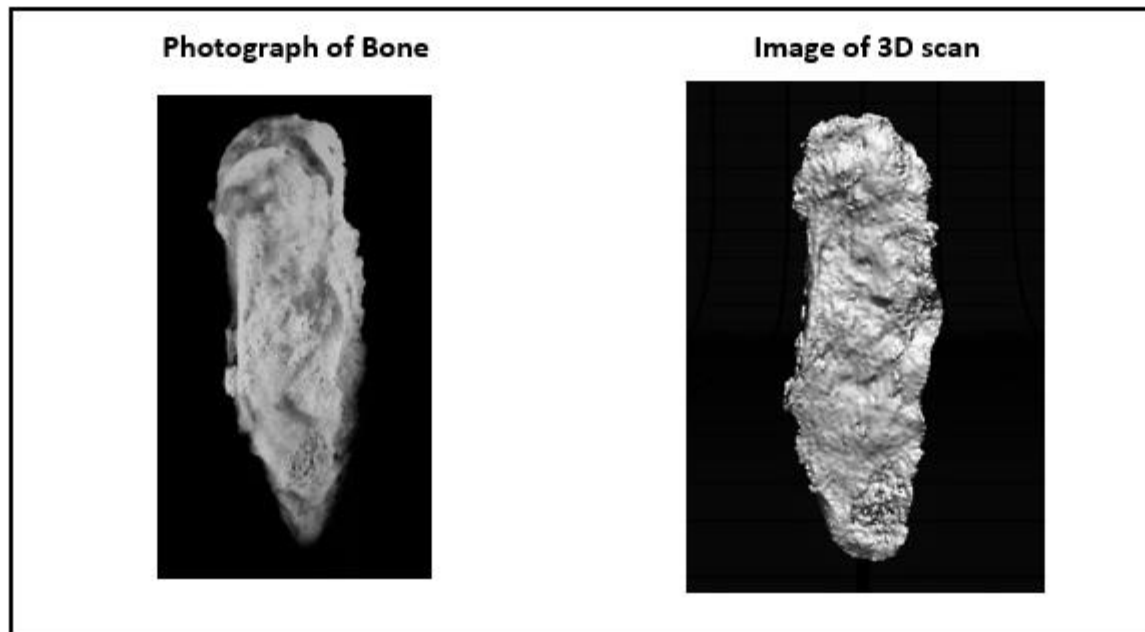


Figure 4.14 Individual 15: Images of Pubic Symphyseal Face

I15 has facial morphology more demonstrative of the features typically associated with advanced age. There is lipping along the dorsal margin, the rim is eroding along the superior ventral margin, there are prominent areas of porosity, and the overall shape of the face is irregular due to emerging ossification. The higher age given to I15 by FANSIE, which is a fair reflection of their facial morphology, highlights that regardless of accuracy the CNN is recognizing features and learning to correlate them to ages. The unusually young appearing face of I16, relative to actual age, and as corroborated by the Observers, is therefore given an erroneous, but true to type, age by the CNN.

## Broad Considerations

### Age Cohort Classification as a Proxy for Accuracy

In order to provide some generalizable metric for evaluating FANSIE's success, individuals in the test set were categorized into Young Adult, Middle Adult, and Old Adult

cohorts. These categories were chosen as they are used in bioarcheology to classify adults in a population, and the nature of the test set lent itself to these divisions. The cohorts were defined (YA 18 - 34, MA 35 - 66, OA 67+) according to the limits of the restricted categories created to minimize phase overlap. Ultimately, however, because the population is of known age, and FANSIE's output is a number not a range, there are potential consequences to putting a firm limit on placing an individual into one cohort or the other, or deeming FANSIE accurate or not based on those limits.

For instance, I chose to define the upper limit of the MA cohort at 66. In archaeological populations, 66 years old would surely be reckoned as an older adult. Here, because the population is comprised of modern Americans, it seemed appropriate. It was only due to the accident of random sampling that awkwardness as a side effect of hard age limits was avoided. For instance, while the MA cohort ends at 66, and the OA cohort begins at 67, the oldest MA Individual is only 62 and the youngest OA Individuals is 68. Further, while the YA cohort ends at 34, its oldest member is only 28; the youngest MA could potentially be 35, but was actually 39. Had the randomized process of choosing the test set yielded both a 34 and 35 year old, the strategy of evaluating accuracy by cohort would be questionable. Deeming the former a 'young adult' and the latter 'middle aged' would create a fine means for altering interpretation of results based on what is ultimately an arbitrary cut-off.

To further complicate the matter, the strategy of defining cohorts was also chosen to provide some guidance in evaluating the accuracy of FANSIE's output relative to the accuracy of Inter-observer classifications. The same problem arises, which is that the cohorts have firm delimitations, while phases overlap. In a number of examples, observers are accurate in their classification, while FANSIE is considered inaccurate, despite producing an age-at-death estimate that falls within the range of the phases assigned by observers. Privileging human accuracy over the machine is considered the correct approach because human osteologists ideally understand that phases are *morphological patterns* that have been demonstrated to correspond to a range of ages, rather than an age range in and of itself. FANSIE, however, was not created to classify morphology, per se, but to classify age. As noted earlier in the discussion, there are currently no disciplinary standards for acceptable accuracy parameters in machine learning. Devising some strategy for holistically assessing machine learning outputs and defining the

parameters of accuracy will be an important aspect of future research. The strategy employed herein is therefore presented with the caveats discussed above.

### Among Young Adults

FANSIE failed to accurately classify any individual in the young adult cohort, while observers were overall correct in their assignments. The notable exception concerns I02, discussed in depth as a case study. I01 presents another interesting case, as humans were able to accurately extrapolate a proper phase despite bony damage, while FANSIE is not capable of such thinking. Based on these results, humans are clearly superior to FANSIE in producing accurate age estimations. Note that FANSIE's age estimate fell within the phase boundaries of I04, who was considered accurately classified by observers. Notably, the age determination produced (49.6) is at the high end of the phase 4 range chosen by two observers, but does not fall within the boundaries of Phase 3, as chosen by Obs 3. So, even here in a potentially ambiguous case of CNN accuracy, we can conclude that humans may produce more accurate results.

It should also be considered that the early phases (1 and 2) have narrow ranges, so classifying an individual into Phase 1, as in I01, produces a neat age range of 15 - 23. Humans are therefore not only better able to classify the young, but are also able to produce a nicely restricted age estimate for young adults with the characteristic morphology of youth. Of course, should a young adult have uncharacteristic morphology, as in I02, then humans are on par with FANSIE, in that both do a bad job of it. Accounting for morphological outliers seems to be a task that both humans and CNNs are unable to reckon with, though a human osteologist will presumably have the opportunity to assess other features of ageing while FANSIE will not. Had the Observers the opportunity to examine a complete skeleton, rather than just the PS face, and been asked to produce the most restricted age determination skeletal evidence supported, then it's very likely they would have noted other important features of the 21 year old I02 and given a younger age estimate, while FANSIE is restricted to the PS face only. Important features such as epiphyseal closure, dentition, and the sternal 4th rib end cannot be considered by FANSIE as they may by humans.

It is unfortunate that a fluke of the random sampling, and the decision to hold firm to that (rather than cherry pick a test set), didn't allow FANSIE to attempt classification of a young adult



who fully displays the classic deep, billowing morphology. It is unknown how the system would assess such an individual. The morphological patterning is easily the most distinctive, but also among the least represented in the training set. Indeed, 'young adults' account for less than 20% of the training set, a factor which no doubt plays into FANSIE's inadequacies in classifying them. Given that youthfulness is both distinct in morphology, while also being poorly represented among the training set, bolstering the batch with more young adult samples would certainly benefit performance. However, because humans do a fine job of classifying these individuals, and have advantages that CNNs will not in the case of outliers, overall CNNs are probably not going to be of much assistance to osteologists in producing accurate age estimations for young adults.

#### Among Middle Adults

The Middle Adult Cohort is where FANSIE's star shines. For individuals in this age category, the CNN is able to produce very accurate age estimations. Further, while Observers were also overwhelmingly correct in their classifications, it should be noted that the phases assigned for middle adults (4, 5, and 6) are all very broad, spanning decades, while FANSIE was accurate within 10 years for all tested Individuals, and within 5 years for 5 of the 7. The CNN is marginally better at classifying than humans overall (100% accuracy, compared to 95.2% accuracy for humans), but excels at producing age estimates that are remarkably close to the real age of the individual. The challenge here is to unpack how meaningful these results may be regarded as, and to address why there is such pronounced accuracy among the middle adults compared to the young and the old.

While a wide range of ages were represented in the training set, FANSIE only produced age ranges between 34 - 60 years. Figure 4.16 demonstrates the trends in age estimation. There are a number of factors which may have contributed to this result. First, bias in the training set must be considered. The tendency of the CNN to produce a middling range of ages runs parallel to the tendency of conventional methods of age estimation to do likewise, a phenomenon dubbed "the attraction of the middle" (Masset, pg. 81, 1989). This has been attributed to age distribution biases in the reference populations. It is certainly possible that such a bias has had an influence on FANSIE - the middle age cohort, the only defined cohort to which any test sample was

assigned, accounts for over half of the training set. There is a type of recursion that is at play. In that FANSIE is best at recognizing the morphologies of the middle ages, the cohort the network is most trained upon, it is likewise prone to assigning a middle age value. Noteworthy features are likely more frequently represented among the middle aged simply due to their overrepresentation within the data set.

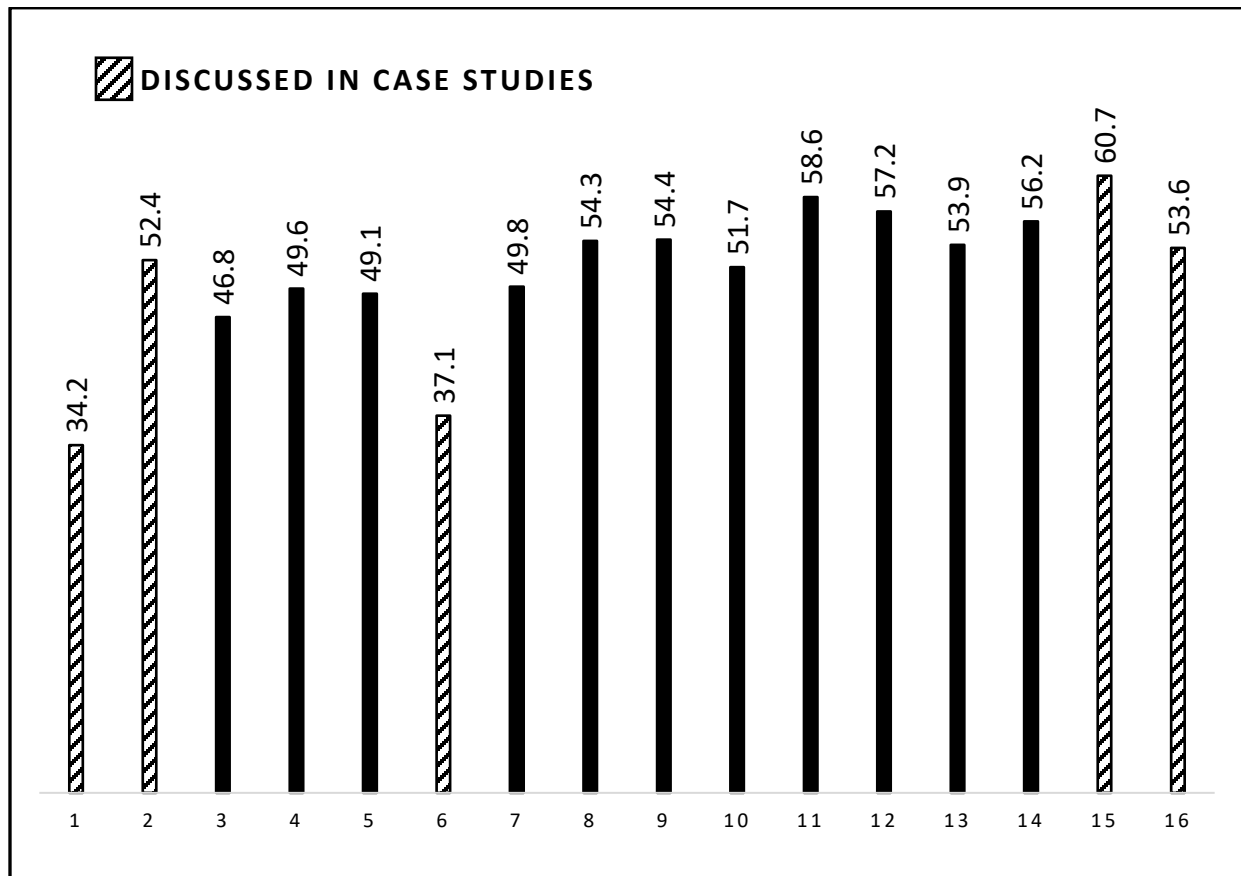


Figure 4.15 CNN Age Output Trends

Further, because the middle aged represent the transition between 'young' morphology and 'old' morphology, there is the assurance that due to variations in rates of bony degeneration, that any particular morphology is likely among them. Thus, the older young and the younger old are both found in the middle. This is, of course, the same phenomenon that contributes to the decades wide spans of the middling SB phases.

It is doubly disappointing that FANSIE's training was unable to be tested on a typical very young adult, the only morphological patterning that is truly restricted to a particularly narrow population. Even the morphologies of the elderly may present at relatively young ages. Evaluating FANSIE in this fashion will be a chief priority in the future. Being unable to readily produce assurance of the system's learning and accuracy using a sure categorization necessitates the more tedious process of teasing out the same with more subtle results. Inter-observer scoring provided an unintended boon in this process, as it afforded a means of explicating minor shifts in output that demonstrated CNN learning. As discussed in a number of case studies presented above, evidence of FANSIE's training is best found in those circumstances where both the network and the Observers share in their misclassification of Individuals. Although FANSIE *only* produced MA ages for the young adults, the classification of I02 is still notably greater than others in that category. Likewise, the Observers also classified I02 as substantially older than other young adults. Therefore, it can be presumed that I02 is a morphological outlier; FANSIE recognizes this individual as 'looking' older than other young adults, and responds appropriately by assigning an older age value.

These justifications are necessary when attempting to interpret FANSIE's results for the middle-aged adults because without this evidence of learning, the accuracy here could be dismissed as being *only* the product of coincidence between an 'attraction towards the middle' in results, and the middle ages of the Individuals. The validation testing performed during training confirms that the CNN is learning, but such strategies aren't particularly satisfying to a human's intuition. As previously discussed, a key component of the successful integration of machine learning technology will be to ensure that the user is able to find some meaning in the results. Making sense of FANSIE's high accuracy among the middle ages is essential if the technology is to have any real-world applications.

The challenge is that post hoc examination is a much more useful strategy when sorting out where FANSIE went wrong, rather than where the network got it very right. It is fairly easy to see what features would lead FANSIE to assume that I02 is older, or that I16 is younger. For an example to the contrary, consider I11 who is 59 years old. FANSIE produced an age of 58.6. As far as physical anthropology has determined, there are no features that are inherently '59 years old-ish'! Clearly, however, there is some aspect of FANSIE's training that does. What that feature

is may be impossible to determine, and the results seem too good to be true. It's important to recall that CNNs do not behave randomly. Whatever output is assigned, regardless of accuracy, is the result of decision making.

The good news here is that high accuracy outputs can be trusted. There is something in the data that has led to the deliberate classification of I11 as 58.6, a remarkably accurate result. The bad news is that there is something in the data that has also led to the deliberate classification of I03 as 46.8, a remarkably inaccurate result. So, accurate results may really only be trusted when it can be demonstrated that they are accurate because actual age-at-death is known! Among young adults, it has already been demonstrated that physical anthropologists do just fine without machines. Perhaps then, a researcher may choose to only employ the services of FANSIE, or a system like her, in circumstances where they have already determined a middle-aged adult status and seek only to refine that age with a tool that works best when augmenting expert categorization, rather than replacing it.

Finally, another factor to consider which may have influenced FANSIE's habit of assigning middle age values is the output method chosen, which uses a linear regression function to produce an age determination. A linear regression method was deemed the most appropriate choice as age related morphological variation progresses at different rates and with different patterning among all individuals. In so doing, however, FANSIE seems to have fallen into the statistical trappings that characterize linear regression outputs. Aykroyd *et al.* argue that classic methods of age determination, which rely on linear regression functions to devise their strategies, tend to overage the young, and underage the old, as both a consequence of the factors discussed above, but also as a property of the linear regression itself (1997). They propose alternative statistical means for generating age output based on correlated features that may be worth attempting for FANSIE.

#### Among Older Adults

FANSIE's performance among the OA individuals in the test set was similar to the YA output in that the trend towards middle aged responses ensured a ubiquitous misclassification of individuals. Individuals in the cohort were unanimously underaged. What distinguishes the discussion of accuracy here, compared to the YA cohort, is the tendency among human observers

to do the same. Underestimation of the elderly is a widely known phenomenon among osteologists (Milner and Boldsen, 2012). Observations of the test set were no exception, and inter-observer scoring only achieved 25% accuracy in classification. It is worth noting that while FANSIE is considered to have 0% accuracy due to the hard limits of the defined age cohorts, no age determination generated fell outside of the 95% CI of assigned SB phases. Preference is given to Observers because they are operating with tested methods that have wide acceptance within the discipline, while FANSIE is afforded no such consideration.

There is also the consideration that having known ages allows for evaluation of accuracy, while use 'in the wild' would naturally preclude that. So, the suggestion to apply FANSIE in circumstances where the researcher has determined a middle-aged status as a means of 'fine tuning' the age estimate would be very counterproductive for the elderly as both the researcher and the CNN will most likely inaccurately assume the individual is younger than in reality. The implications of the such, and other factors to reckon with when assessing the potential for machine learning in anthropological work will be examined in 'Chapter V: Conclusions' to follow.

## CHAPTER V

### CONCLUSIONS AND FUTURE DIRECTIONS

#### Applied Utility

At this juncture, the machine learning approach undertaken in this research does not appear to be a useful approach to the problem of accurate age-at-death estimation in anthropology. Recall the discussion of barriers to integration presented in Chapter IV. The CNN is prone to errors in cognition where humans do not make errors (as in I01), and therefore not trustworthy as a matter of perspective among likely users seeking to refine age estimation of young adults. Mistrust of FANSIE's results is appropriate - humans do a fine job of assessing age without machine assistance, and indeed outperform the machine. Among younger individuals, the traditional pubic symphysis ageing methods perform better than the CNN. The features of the young are distinct morphologically and thus unlikely to be misclassified by a skilled osteologist, except, of course, in those circumstances of aberrant morphology wherein FANSIE is susceptible to the same challenges.

Among the elderly, FANSIE's performance is at best on par with human observers. FANSIE also seems to be prone to the same kinds of errors that humans make when determining age in the elderly. Among potential users the CNN may thus be considered more trustworthy, but as its categorizations are no better than what a human produces, they are not particularly useful. Conversely, FANSIE's abilities in determining age at death among the middle aged are probably *too* accurate to be regarded with anything other than suspicion by a user. Further, while FANSIE is capable of producing remarkably accurate responses among the middle aged, those responses can only be established as such with the benefit of known demography. The results are intriguing from a purely academic perspective, but won't do much good in the applied contexts for which FANSIE was designed. Even if a researcher were to use FANSIE to fine-tune their age at death estimation of a presumed middle-aged individual, they are likely to erroneously assume that an

older adult is middle-aged. FANSIE will do the same and subsequently reify a mistaken age assumption, which is perhaps worse than doing nothing at all. Using the CNN in this manner may be further challenged as presumably these methods will be restricted to bioarcheologists, rather than forensic anthropologists, and there are potential incompatibilities between the morphologies of archaeological populations and the modern population FANSIE is trained upon.

At this point, FANSIE's results cannot be trusted as either a matter of actual accuracy, vis-à-vis the real age of the individual, or as a matter of human faith in machines vis-à-vis successful integration into practice. It was considered during early project planning that rather than having FANSIE generate outputs of distinct ages, the CNN could instead be programmed to assign test subjects to broader categories, and the results so far indicate that this strategy would yield better overall accuracy, if reduced precision. To avoid the problems resulting from hard cut-offs, overlapping categories would be created and training subjects would be identified according to all appropriate categories. For example, a 35-year-old training subject would be identified as fitting into both a hypothetical 27 - 37 category, and a 35 - 45 category.

That idea was abandoned when it was quickly realized that in so doing, a machine would be created to do exactly what human researchers had *already done*, and was exactly the sort of thing that we were working to replace. However, it may be that due to individual morphological expression, and rates of skeletal transformations, that these sorts of strategies are really the best that are possible when assessing gross anatomy. Thus far, the machine learning approach employed here does not outperform humans using human-devised methods, and may not ever do so. However, there are some confounding factors to consider that, once controlled, may improve FANSIE's capabilities.

## Project Limitations

### Training Sample Size

Any discussion of CNN accuracy must consider the effects of training sample size on output. There is no inherent 'ideal' training sample size, rather the volume of training data required to do the job is contingent upon the task itself. However, it has been demonstrated that CNNs trained with larger batches of data outperform identical networks with smaller training

batch sizes (Radiuk, 2017). The training sample of  $n=292$ , is a fairly robust sample size for most anthropological research, but in the context of machine learning it is rather small. This is particularly so given the nature of the feature being categorized. There are aspects of the pubic symphysis captured in the scanning that are considered irrelevant to age estimation, such as the height and width of the feature. These kinds of details are no doubt factoring into decision making, so a larger training batch will ultimately be necessary to better assist FANSIE in separating the wheat from the chaff.

An analogous study was performed by Esteva *et al.* (2017) wherein they trained a CNN to classify various skin lesions based on photographs. In that study, the CNN was able to outperform board-certified dermatologists in lesion categorization. I regard this work as analogous because like the pubic symphysis, skin lesions have a variety of morphological aspects, such as relative size, that are either less relevant or wholly irrelevant to classification compared to more salient features. Further, there are only moderate differences between the look of each lesion, much like in the pubic symphysis.

Recognizing which features are more or less diagnostic, and learning the importance of subtle morphological variations is a big task and certainly will require a lot of education. Esteva and colleagues trained their CNN on 12,000 images, but a training sample of that size is not realistic for anthropology. Beyond the time burden necessary to scan that many pubic symphyses, there are the limitations associated with accessing that many skeletal remains of modern individuals with certain demography. Even if high accuracy is hypothetically possible, actually achieving it may not be due to the impossibility of sourcing an adequate number of appropriate specimens for the training set.

## Scanning Technique

A recurring challenge encountered during the data collection involved mitigating the complications resulting from the laser light based approach to digital rendering. The contours of the bone itself made capturing the entirety of the pubic symphysis, including the crucial side context, somewhat difficult. As previously discussed, the straight line of the laser can only register the presence of bony material if it impacts the feature head-on. In order to fully capture all of the dorsal side context, where the face meets the bone in a convex arc, a much more time



consuming process of 180-degree scanning would be necessitated. To be clear, not every scan is lacking a complete dorsal side aspect, but it was a common enough error that there are a number of scans with incomplete, marginal, or absent dorsal side context. Some of the features identified by osteologists as being important to age determination are associated with activity along the dorsal border, so the absence of data in that region could be an impediment to FANSIE's ability to learn essential relationships.

Another aspect of the scanning technique that may be limiting proper representation of the feature was the decision to maximize image resolution. The highest quality scanning technique protocol was employed, with the intention of capturing as much minute data as possible. Unfortunately, this well-intentioned decision may have been a hindrance rather than a help. The light point density of maximum resolution is too sensitive for the chore. In some cases, tiny flaws in the face were amplified with this protocol. Figure 5.01 shows a scan of the pubic symphyseal face taken with both high definition and low definition scanning protocols. The microscopic porosities in the face are amplified in the macro scale, presumably because the distance between 'hits' - laser light impacts to the bone, and 'misses' are too widely spaced at this high resolution to produce vertex density sufficient to display as solid bone. Conversely, the low-density technique produces a much more faithful image.

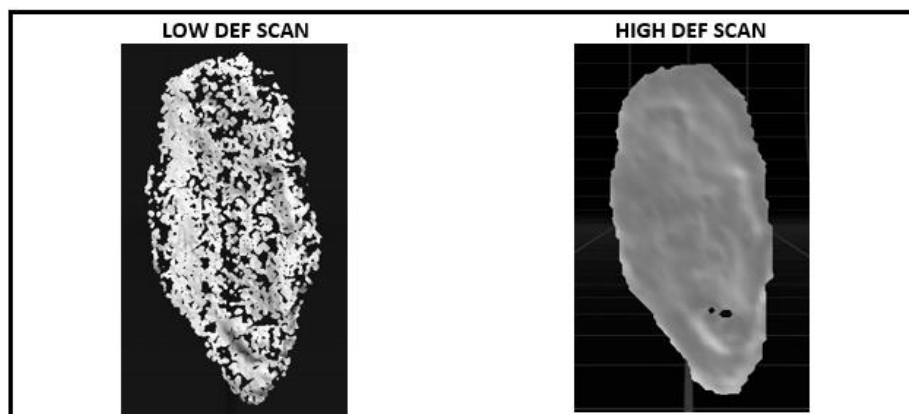


Figure 5.1 High Def and Low Def Scan Techniques

Unfortunately, this wasn't discovered until after the data collection process. At the time, it was presumed that maximizing vertex density would produce the best high fidelity digital representation, but it seems that this assumption was erroneous. A further complication is that in

some scenarios the higher resolution technique probably is the best approach, as the low-density technique will miss important features like small macroporosities and the like. It may be that a variety of scanning protocols are indicated, depending upon the nature of the bone, to produce the best representation of the feature. Low definition scanning is substantially quicker than high definition, so in the future starting with lower resolution and working up to capture necessary details as needed will be the more appropriate choice.

### Computational Variables

CNNs and other machine learning applications are experiencing a renaissance of research, but thus far the ideal specifications for particular tasks are determined heuristically on a case by case basis. Ideal vary because the computational methods that give CNNs their predictive power may rather flexibly executed. For example, FANSIE was designed with 20 deep layers which convolve and otherwise modify the input data in particular ways before determining output, but the 20-layer construction is essentially arbitrary. FANSIE could have just as readily been designed with 10 layers, or 30. It is certainly possible that increasing data filtration through more hidden layers may prove useful in increasing accuracy, but at this point that is undetermined. It may also be that the data is overly processed, and would benefit from decreased filtration.

The ideal parameters for any particular project, included FANSIE, are typically fine-tuned manually through modification of the source code before settling on the optimal arrangement of numbers and types of layers best suited to the job. The computational costs of managing such a chore represent another important challenge. While a more complex CNN *may* produce superior results, increasingly complex systems require increasingly greater computational power to train. The computational demands of training a modest 20-layer CNN like FANSIE will increase quite a bit once the training set becomes large enough.

Coupling a large training set to a deeper system will increase those demands exponentially, and eventually necessitate the use of a supercomputer or computational cluster. This is not an insurmountable roadblock to progress on the project, per se, but will have to reckoned with as research continues, and has already been a deciding factor in protocols. Most importantly, the choice to voxelize the PS scans on a 38x38x38 grid was a concession to

managing FANSIE's computational needs. The 38x38x38 grid does a fair job of representing the image, but does flatten out features that might otherwise be diagnostically useful. However, increasing the voxel density, as in a 64x64x64 grid, would require substantially more computing power to process. Reconciling a useful balance between the exponential computational demands that come with a deep system operating on high-volume, complex data with the ideal parameters to achieve high accuracy will require trial-and-error troubleshooting as research continues.

### FANSIE in the Future

While FANSIE is not yet operating at a level that will afford integration into practice, there are number of important factors to address before writing off machine learning in anthropology all together. As discussed in 'Project Limitations', increasing the training sample set will be the most crucial factor in further assessment of FANSIE's potential. The data set already produced is a good start, but is far too small to fully exploit the potential of machine learning. Cho *et al.* (2015) have developed a learning curve approach to extrapolate the ideal training sample size using a function based assessment of outputs with different training batch sizes. Applying this technique to FANSIE will help establish if her data needs are feasible, and provide a goal to work towards.

There are several other factors that could be modified or improved upon to promote future performance beyond increasing the data set, and have the bonus of only requiring the coding hours necessary to perform them. Should increasing the training set prove to be an unrealistic goal either due to the impossibility of the number required, or the inability to procure the scan data, then tinkering with FANSIE's code will be the next step. Giving FANSIE a fair opportunity to prove herself before dismissing the usefulness of machine learning in anthropology altogether is essential; machine learning has already been successfully applied to a number of novel and unanticipated contexts. A more thorough assessment of the technology is indicated before anything can be said with certainty.

First, there is the option to employ the statistical output method advised by Akyroyd *et al.* (1997), as discussed in Chapter IV, and abandon the linear regression function output method currently in use. It may also be considered that the 'age-range category' output strategy is not such a poor idea altogether if success can be had in narrowing the margins of those ranges

beyond those currently employed. A potential strategy could involve sorting into a series of progressively tighter ranges based on discriminate function analysis of best fit. Or, a direct approach could be tried, eschewing statistics based outputs altogether, and instead simply directing FANSIE just pick a number between 18 - 100 that seems best. Each of these strategies will, by function, produce different results and it may be that one of them will do so more accurately than the current method in use.

There is also the option to increase the data set with scans that are already available by including female subjects and right sided scans, which would yield a training batch of over 700, compared to the  $n=292$  used here. Of course, it must be considered that the consequences of introducing either right sided scans or female individuals could just as readily decrease accuracy. FANSIE's skill in assessing I06, a right aspect scan, indicates that sidedness may not be a crucial factor in assessment. It has been suggested that female skeletal morphology transforms in less predictable ways than that of males (Hartnett, 2010), but the benefits conferred by batch increase may outweigh those costs. Employing either method alone warrants implementation if only to assess the output. Should performance increase or decrease in each case, then that in and of itself is noteworthy.

CNNs may learn to make associations that are dismissed as irrelevant by humans, but still produce good results. It was stated previously that FANSIE is no doubt making correlations between aspects of the data, such as height and width of the feature, that are regarded as irrelevant by osteologists, but still influential to her decision making. However, just because the dimensions of the PS face are considered irrelevant, it doesn't necessarily mean that they actually are. Given that wear patterning of the face is inherently dependent upon activity at the site of the joint, it is fair to speculate that variations in surface area might play some role in determining the pace and nature of those changes.

Anecdotally, a number of face 'types' were observed during data collection, and the frequency and fidelity of those types was such that it was surprising that, as far as I am aware, these variations have not been commented upon. Some of the bony context that informs the character of a face type are considered when making sex determinations, such as the robusticity of the ramii, but as yet are not regarded as meaningful in the context of age determination. Making these sorts of associations, which seem counterintuitive or nonproductive to humans, are

where CNNs excel, and are the source of the greatest potential to be an asset to human researchers.

The ultimate advantage of the machine is exactly that which has long been considered its greatest weakness - machine systems are not capable of innately biased cognition, what might be more generously called 'intuition'. When tasked to analyze an input, they will consider every facet of that input equally with no *a priori* assumptions. Efficiently assessing a huge number of correlates for a large volume of data is exactly the sort of work that computers are much better suited to perform than humans. Sometimes this goes comically awry, as many famous cases can testify to, but sometimes the machine gets it right in ways that would never have occurred to a human, but which may be identified and exploited by them. The ultimate hope for FANSIE is not that the network will replace humans, but that it will better assist them, by applying a new way of thinking to an old problem.

## WORKS CITED

- Aykroyd, R., Lucy, D., Pollard, A., & Solheim, T. (1997). Technical note: Regression analysis in adult age estimation. *American Journal of Physical Anthropology*, 104(2), 259-265.
- Biwasaka, Sato, Aoki, Kato, Maeno, Tanijiri, . . . Dewa. (2013). Three dimensional surface analyses of pubic symphyseal faces of contemporary Japanese reconstructed with 3D digitized scanner. *Legal Medicine*, 15(5), 264-268.
- Brooks, S., & Suchey, J. M. (1990). Skeletal age determination based on the os pubis: A comparison of the Acsádi-Nemeskéri and Suchey-Brooks methods. *Human Evolution*, 5(3), 227-238. doi:10.1007/bf02437238
- Burrell, J. (2015). How the Machine Thinks: Understanding Opacity in Machine Learning Algorithms. *SSRN Electronic Journal*. doi:10.2139/ssrn.2660674
- Caruana, R., Kangaroo, H., Dionisio, J., Sinha, U., & Johnson, D. (1999). Case-based explanation of non-case-based learning methods. *Proceedings. AMIA Symposium*, 212-5.
- Daubert v. Merrell Dow Pharmaceuticals, Inc. (1993) 509 U.S. 579
- Dieleman, S., Willett, K., & Dambre, J. (2015). Rotation-invariant convolutional neural networks for galaxy morphology prediction. *Monthly Notices of the Royal Astronomical Society*, 450(2), 1441-1459.
- Esteva, A., Kuprel, B., Novoa, R., Ko, J., Swetter, S., Blau, H., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639), 115-118G.
- Frye v. United States. (1923). 293 F. 1013 D.C. App.
- Hartnett, K. (2010). Analysis of Age-at-Death Estimation Using Data from a New, Modern Autopsy Sample - Part I: Pubic Bone. *Journal of Forensic Sciences*, 55(5), 1145.
- Junghwan Cho, Kyewook Lee, Ellie Shin, Garry Choy, and Synho Do. (2015). How much data is needed to train a medical image deep learning system to achieve necessary high accuracy? doi:atXiv:1511.06348
- Lee I, Martin F, Denner J, et al. (2011). Computational thinking for youth in practice. *ACM*

- Inroads 2(1): 32–37.
- Lipton, Z.C. (2017). The mythos of model interpretability. doi:arXiv:1606.03490
- Masset C. (1989). Age estimation on the basis of cranial sutures. In MY Iscan (ed.): *Age Markers in the Human Skeleton*. Springfield: C.C. Thomas, pp. 71-103.
- Milner, G., & Boldsen, J. (2012). Transition analysis: A validation study with known-age modern American skeletons. *American Journal of Physical Anthropology*, 148(1), 98-110.
- Radiuk, P. (n.d.). Impact of Training Set Batch Size on the Performance of Convolutional Neural Networks for Diverse Datasets. *Information Technology and Management Science*, 20(1), 20-24.
- Ritz-Timme, S., Cattaneo, C., Collins, M., Waite, E., Schütz, H., Kaatsch, H., & Borrman, H. (2000). Age estimation: The state of the art in relation to the specific demands of forensic practise. *International Journal of Legal Medicine*, 113(3), 129-36.
- Slice, D., & Algee-Hewitt, B. (2015). Modeling Bone Surface Morphology: A Fully Quantitative Method for Age-at-Death Estimation Using the Pubic Symphysis. *Journal of Forensic Sciences*, 60(4), 835-843.
- Stoyanova, D., Algee-Hewitt, B., & Slice, D. (2015). An enhanced computational method for age-at-death estimation based on the pubic symphysis using 3D laser scans and thin plate splines. *American Journal of Physical Anthropology*, 158(3), 431-440.
- Todd, T. (1920). Age changes in the pubic bone. I. The male white pubis. *American Journal of Physical Anthropology*, 3(3), 285-334.
- Wing JM (2006) Computational thinking. *Communications of the ACM* 49(3): 33–35.

## APPENDIX A

### Technical Report by James Jenkins

#### CNN Pipeline

Initial step is to process the .obj files storing the following data in a binary representation to reduce data size.

- Scan ID
- Age
- Gender
- Bracket ID
- N verticies
- 64bit floating point (double) triples for every vertex

Once this level of processing has been completed, the file is loaded and rendered as a voxelized version. 38x38x38 dimensions are used to represent the file. Larger voxelization parameters would provide better accuracy, but storage and computation time rapidly exceeds reasonable parameters. The voxelizations are the original scan rotated about each of the major x,y,z axes in 5 degree increments then voxelized into a 38x38x38 space. The rotation followed by voxelization provides a means of dataset augmentation. This is especially useful, given the relatively small size of the dataset.

CNN models were trained with the following layers:

- Convolution 20 filters, kernel 5, stride 1
- Max pool
- Convolution 10 filters, kernel 5, stride 1
- Dense layer 1000



- Dense layer 100
- Dense layer 10
- Single output with no activation for regression

Final results were generated by bagging with 3 of the models.

## APPENDIX B

### Results by Individual

#### Individual 01

Individual 01 has a known age-at-death of 18 years. Individual 01 is classified as a young adult. FANSIE determined the age of Individual 01 to be 34.2, overestimating the known age by 16.2 years, and exceeding the upper limit of the YA range by 0.2 years. All Observers were in agreement and accurately placed Individual 01 into SBP 1.

Individual 01			
Known Age	CNN Age		CNN Error
18	34.2		+ 16.2
Intraobserver Classifications (Suchey-Brooks Method)			
Observers	Phase	Mean Age	95% CI
OBS 1	1	18.5	15 - 23
OBS 2	1	18.5	15 - 23
OBS 3	1	18.5	15 - 23

#### Individual 02

Individual 02 has a known age-at-death of 21 years. Individual 02 is classified as a young adult. FANSIE determined the age of Individual 02 to be 52.4, overestimating the known age by 31.4 years, and exceeding the upper limit of the YA age range by 31.4 years. Observers were in

disagreement. Observers 1 and 3 overestimated the age of this individual by assigning an SBP of 4. Observer 2 also overestimated the age of Individual 02, assigning an SB score of 5.

Individual 02			
Known Age	CNN Age		CNN Error
21	52.4		+ 31.4
Intraobserver Classifications (Suchey-Brooks Method)			
Observers	Phase	Mean Age	95% CI
OBS 1	4	35.2	23 - 57
OBS 2	5	46.6	27 - 66
OBS 3	4	35.2	23 - 57

#### Individual 03

The known age-at-death of Individual 03 is 22 years. Individual 03 is classified as a young adult. FANSIE determined the age of Individual 03 to be 46.8, overestimating the known age by 24.8 years, and exceeding the upper limit of the YA age range by 24.8 years. All observers were in agreement and accurately classified Individual 03 with an SBP score of 2.

Individual 03			
Known Age	CNN Age		CNN Error
22	46.8		+ 24.8
Intraobserver Classifications (Suchey-Brooks Method)			
Observers	Phase	Mean Age	95% CI
OBS 1	2	23.4	19 - 34
OBS 2	2	23.4	19 - 34
OBS 3	2	23.4	19 - 34

#### Individual 04

The known age-at-death of Individual 04 is 27 years. Individual 04 is classified as a young adult. FANSIE determined the age of Individual 04 to be 49.6, overestimating the age by 22.6 years, and exceeding the upper limit of the YA age range by 22.6 years. Observers were in disagreement. Observers 1 and 2 provided an SBP score of 4, while Observer 3 assigned a score of 3.

Individual 04			
Known Age	CNN Age		CNN Error
27	49.6		+ 22.6
Intraobserver Classifications (Suchey-Brooks Method)			
Observers	Phase	Mean Age	95% CI
OBS 1	4	35.2	23 - 57
OBS 2	4	35.2	23 - 57
OBS 3	3	28.7	21 - 46

#### Individual 05

Individual 05 has a known age-at-death of 28 years. Individual 04 is classified as a young adult. FANSIE determined the age of Individual 05 to be 49.1, overestimating the age of this individual by 21.2 years, and exceeding the upper limit of the YA age range by 15.1 years. All Observers were in agreement, and accurately placed this individual in SBP 3.

Individual 05			
Known Age		CNN Age	CNN Error
28		49.1	+ 21.2
Intraobserver Classifications (Suchey-Brooks Method)			
Observers	Phase	Mean Age	95% CI
OBS 1	3	28.7	21 - 46
OBS 2	3	28.7	21 - 46
OBS 3	3	28.7	21 - 46

#### Individual 06

Individual 06 has a known age-at-death of 39 years. Individual 06 is classified as a middle adult. FANSIE determined the age of Individual 06 to be 37 years, and underestimated the age of this individual by 2 years. All Observers were in agreement, and correctly placed Individual 06 within SBP 4.

Individual 06			
Known Age	CNN Age		CNN Error
39	37		- 2
Intraobserver Classifications (Suchey-Brooks Method)			
Observers	Phase	Mean Age	95% CI
OBS 1	4	35.2	23 - 57
OBS 2	4	35.2	23 - 57
OBS 3	4	35.2	23 - 57

## Individual 07

Individual 07 has a known age-at-death of 40 years. Individual 07 is classified as a middle adult. FANSIE determined the age of Individual 07 to be 49.8, overestimating the age by 9.8 years. Observers were in disagreement. Observers 1 and 3 placed Individual 07 into SBP 5; Observer 2 placed Individual 07 into SBP 4.

Individual 07			
Known Age	CNN Age		CNN Error
40	49.8		+ 9.8
Intraobserver Classifications (Suchey-Brooks Method)			
Observers	Phase	Mean Age	95% CI
OBS 1	5	46.6	27 - 66
OBS 2	4	35.2	23 - 57
OBS 3	5	46.6	27 - 66

## Individual 08

Individual 08 has a known age-at-death of 47 years. Individual 08 is classified as a middle adult. FANSIE determined the age of Individual 08 to be 54.3, overestimating the age by 7.3 years. Observers were in agreement, and correctly placed Individual 08 within SBP 4.

Individual 08			
Known Age	CNN Age		CNN Error
47	54.3		+ 7.3
Intraobserver Classifications (Suchey-Brooks Method)			
Observers	Phase	Mean Age	95% CI
OBS 1	4	35.2	23 - 57
OBS 2	4	35.2	23 - 57
OBS 3	4	35.2	23 - 57

#### Individual 09

Individual 09 has a known age-at-death of 50 years. Individual 09 is classified as a middle adult. FANSIE determined the age of Individual 09 to be 54.4, overestimating the age by 4.4 years. Observers were in disagreement. Observers 1 and 3 placed Individual 09 into SBP 5, while Observer 2 assigned an SBP of 4.

Individual 09			
Known Age	CNN Age		CNN Error
50	54.4		+ 4.4
Intraobserver Classifications (Suchey-Brooks Method)			
Observers	Phase	Mean Age	95% CI
OBS 1	5	46.6	27 - 66
OBS 2	4	35.2	23 - 57
OBS 3	5	46.6	27 - 66

## Individual 10

Individual 10 has a known age-at-death of 57 years. Individual 10 is classified as a middle adult. FANSIE determined the age of Individual 10 to be 51.8 years, underestimating the age by 5.2 years. Observers were in disagreement. Observer 1 placed Individual 10 in SBP 6, while Observers 2 and 3 assigned a score of 5.

Individual 10			
Known Age	CNN Age		CNN Error
57	51.8		- 5.2
Intraobserver Classifications (Suchey-Brooks Method)			
Observers	Phase	Mean Age	95% CI
OBS 1	6	61.2	34 - 86
OBS 2	5	46.6	27 - 66
OBS 3	5	46.6	27 - 66

## Individual 11

Individual 11 has a known age-at-death of 59 years. Individual 11 is classified as a middle adult. FANSIE determined the age of Individual 11 to be 58.6 years, underestimating the age by 0.4 years. Observers were in disagreement. Observers 1 and 3 placed Individual 11 into SBP 5. Observer 2 assigned a SBP score of 4. Observers 1 and 3 were correct in their classification; Observer 2 is incorrect.



Individual 11			
Known Age	CNN Age		CNN Error
59	58.6		- 0.4
Intraobserver Classifications (Suchey-Brooks Method)			
Observers	Phase	Mean Age	95% CI
OBS 1	5	46.6	27 - 66
OBS 2	4	35.2	23 - 57
OBS 3	5	46.6	27 - 66

## Individual 12

Individual 12 has a known age-at-death of 62 years. Individual 12 is classified as a middle adult. FANSIE determined the age of Individual 12 to be 57.2, underestimating the age by 4.8 years. Observers were in disagreement. Observer 1 placed Individual 12 into SBP 6, while Observers 2 and 3 assigned an SBP score of 5.

Individual 12			
Known Age	CNN Age		CNN Error
62	57.2		- 4.8
Intraobserver Classifications (Suchey-Brooks Method)			
Observers	Phase	Mean Age	95% CI
OBS 1	6	61.2	34 - 86
OBS 2	5	46.6	27 - 66
OBS 3	5	46.6	27 - 66

### Individual 13

Individual 13 has a known age-at-death of 68 years. Individual 13 is classified as an old adult. FANSIE determined the age of Individual 13 to be 53.9, underestimating the age by 14.1 years. Observers were in disagreement. Observers 1 and 3 placed Individual 13 within SBP 5, while Observer 2 assigned an SBP of 4. All Observers underestimated the age of Individual 13

Individual 13			
Known Age	CNN Age		CNN Error
68	53.9		- 14.1
Intraobserver Classifications (Suchey-Brooks Method)			
Observers	Phase	Mean Age	95% CI
OBS 1	5	46.6	27 - 66
OBS 2	4	35.2	23 - 57
OBS 3	5	46.6	27 - 66

### Individual 14

Individual 14 has a known age-at-death of 68 years. Individual 14 is classified as an old adult. FANSIE determined the age of Individual 14 to be 56.2, underestimating the age by 11.8 years. Observers were in disagreement. Observers 1 and 2 placed Individual 14 within SBP 6, while Observer 3 assigned an SBP of 5. Observers 1 and 2 are considered correct, while Observer 3 underestimated the age and is incorrect.

Individual 14			
Known Age	CNN Age		CNN Error
68	56.2		- 11.8
Intraobserver Classifications (Suchey-Brooks Method)			
Observers	Phase	Mean Age	95% CI
OBS 1	6	61.2	34 - 86
OBS 2	6	61.2	34 - 86
OBS 3	5	46.6	27 - 66

#### Individual 15

Individual 15 has a known age-at-death of 82 years. Individual 15 is classified as an old adult. FANSIE determined the age of Individual 15 to be 60.7 years, underestimating the age by 21.3 years. Observers were in disagreement. Observer 1 correctly placed Individual 15 within SBP 6. Observers 2 and 3 underestimated the age of this individual and assigned an SBP of 5.

Individual 15			
Known Age	CNN Age		CNN Error
82	60.7		- 21.3
Intraobserver Classifications (Suchey-Brooks Method)			
Observers	Phase	Mean Age	95% CI
OBS 1	6	61.2	34 - 86
OBS 2	5	46.6	27 - 66
OBS 3	5	46.6	27 - 66

## Individual 16

Individual 16 has a known age-at-death of 91 years. Individual 16 is classified as an old adult. FANSIE determined the age of Individual 16 to be 53.6, underestimating the age by 31.4 years. Observers were in disagreement. Observer 1 placed Individual 16 within SBP 6. Observer 2 and 3 underestimated the age of Individual 16, assigning an SBP of 5.

Individual 16			
Known Age	CNN Age		CNN Error
91	53.6		- 37.4
Intraobserver Classifications (Suchey-Brooks Method)			
Observers	Phase	Mean Age	95% CI
OBS 1	6	61.2	34 - 86
OBS 2	5	46.6	27 - 66
OBS 3	5	46.6	27 - 66