

Applications of Machine Learning in Forensic Investigations of Human Decomposition

A Dissertation Presented for the
Doctor of Philosophy
Degree
The University of Tennessee, Knoxville

Anna-Maria Nau

December 2024

© by Anna-Maria Nau, 2024
All Rights Reserved.

This work is dedicated to my parents, I could not have done this without them.

Thank you for all your support and love along the way.

Acknowledgements

First and foremost, I would like to thank my advisor, Dr. Audris Mockus, for his guidance, support, and encouragement throughout my Ph.D. journey. His expertise and boundless knowledge have been instrumental in shaping both my dissertation and my growth as a researcher and data scientist. It has been a true pleasure to work with him.

I would also like to thank the members of my committee, Dr. Russell Zaretzki, Dr. Dawnie Wolfe Steadman, and Dr. Amir Sadovnik for their valuable support and feedback throughout my doctoral research.

In particular, I would like to thank Dr. Dawnie Steadman, my research partner Phillip Ditto, and my colleagues from the Anthropology department for their collaboration. They have taught me a great deal about human decomposition and forensic anthropology, which has been very influential throughout my work.

Lastly, I would like to thank my parents, Petra and Thomas, for their unconditional support and encouragement throughout these years. Without them, I would not be the person I am today.

Abstract

Determining the postmortem interval (PMI)—the time elapsed between death and the discovery of a body—is a fundamental challenge in forensic anthropology. Accurate PMI estimation is essential for narrowing down the list of potential decedents, facilitating the identification process, and assisting law enforcement in eliminating suspects or verifying witness statements in cases of homicide. Additionally, knowledge of the PMI provides critical insights into the environmental and biological factors that have influenced the decomposition process over time, thereby enabling a more comprehensive analysis of the remains.

The Forensic Anthropology Center (FAC) at the University of Tennessee, Knoxville (UTK), is home to the Anthropology Research Facility (ARF), where human decomposition has been studied since 1981. Since 2012, photographic documentation of approximately 100 donors per year has resulted in a repository of over one million images from more than 800 donors. This extensive photographic collection, presents an unparalleled resource for forensic research.

In 2016, the Departments of Electrical Engineering and Computer Science and Anthropology at UTK embarked on a collaborative project funded by the National Institute of Justice (NIJ). The project’s primary objectives are to make this vast photographic collection accessible and useful for forensic research and casework and to develop advanced artificial intelligence (AI) tools specifically for forensic purposes, with a primary focus being PMI estimation.

This dissertation contributes to these ongoing efforts by: (1) continuing to develop and maintain a collaborative tool designed to organize, visualize, analyze, and curate the large collection of human decomposition photographs, thus enhancing their usability in forensic research; (2) creating a method to label stages of decay when labeled data is limited and subsequently devising an AI-driven approach to automate the identification of these stages; and (3) evaluating existing PMI estimation methods using this extensive dataset, providing a more comprehensive and statistically significant analysis than earlier studies. Through these contributions, this work aims to advance the field of forensic anthropology by leveraging AI and large-scale data to improve the accuracy and reliability of PMI estimation, thereby enhancing the overall effectiveness of forensic investigations.

Table of Contents

1	Introduction	1
1.1	Problem Statement	1
1.2	Background and Literature Review	3
1.3	Research Objectives	5
1.4	Dissertation Structure	6
2	A Collaborative Tool for Human Decomposition Research and Casework	8
2.1	Abstract	9
2.2	Introduction	10
2.3	Materials and Methods	12
2.3.1	Requirement Gathering	12
2.3.2	Design and Implementation	14
2.3.3	System Evaluation	17
2.4	Results and Discussion	19
2.4.1	Requirement Gathering, Design, and Implementation	19
2.4.2	System Evaluation via a User Study	21
2.5	Conclusion	30
3	Stage of Human Decay Classification Given Limited Labeled Data	31
3.1	Abstract	32
3.2	Introduction	32

3.3	Proposed Vision System	34
3.3.1	The Human Decomposition Dataset	34
3.3.2	Manual Data Annotation	36
3.3.3	Proposed Label Propagation Method	39
3.3.4	SOD Classification	41
3.3.5	Implementation and Experimental Settings	41
3.4	Experimental Results and Discussion	42
3.5	Conclusion	45
4	Towards Automation of Human Stage of Decay Identification: An Artificial Intelligence Approach	46
4.1	Abstract	47
4.2	Introduction	48
4.3	Materials and Methods	49
4.3.1	Stage of Decay Scoring Methods	49
4.3.2	The Human Decomposition Dataset	50
4.3.3	Data Processing and Labeling	51
4.3.4	AI Model Development	52
4.3.5	Interrater Test	55
4.4	Results	57
4.4.1	SOD Classification	57
4.4.2	Interrater Test	58
4.4.3	Discussion	59
4.5	Conclusion	61
5	Identifying Factors to Help Improve Existing Decomposition-Based PMI Estimation Methods	63
5.1	Abstract	64
5.2	Introduction	65
5.3	Materials and methods	67

5.3.1	The Human Decomposition Photographic Collection	67
5.3.2	The Study Sample	67
5.3.3	Decomposition Measurement	70
5.3.4	Weather Data and ADD calculation	71
5.3.5	PMI/ADD Estimation	71
5.4	Results	79
5.5	Discussion	83
5.6	Conclusion	86
6	Conclusion	87
6.1	Research Overview	88
6.2	Discussion, Contributions, and Future Work	88
6.3	Conclusion	91
	Bibliography	92
	Appendices	100
A	Chapter 2 Supporting Information	101
A.1	The Complete Nomenclature.	101
A.2	ETL Data Pipeline.	102
B	Chapter 4 Supporting Information	103
B.1	Megyesi et al.'s [33] decomposition scoring method for the head. 103	
B.2	Megyesi et al.'s [33] decomposition scoring method for the torso. 104	
B.3	Megyesi et al.'s [33] decomposition scoring method for the limbs. 105	
B.4	Gelderman et al.'s [17] decomposition scoring method for the head.	106
B.5	Gelderman et al.'s [17] decomposition scoring method for the torso.	107
B.6	Gelderman et al.'s [17] decomposition scoring method for the limbs.	108

C	Chapter 5 Supporting Information	109
Vita		110

List of Tables

2.1	Example keywords from the final nomenclature of forensic-related keywords.	20
3.1	The SHD-3 and SHD-4 datasets along with their DS groups, class names, and class frequencies (N). Note, the class active decay = early + advanced decay.	39
3.2	Performance comparison of the baseline method, LPA [48], and the proposed method, ExoEndo-LP, when applied to the train data (n=3312). Reported are average runtimes (in seconds, s) and the number of additionally labeled (NoL) train samples using label propagation. . .	43
3.3	SOD-VS with the baseline method, LPA [48] (neighbors=5), and the proposed ExoEndo-LP evaluated on the SHD-3 test data (top three) and SHD-4 test data (bottom three).	44
4.1	The original SOD terms mapped to the new SOD class labels used throughout this study for the different scoring methods.	52
4.2	The labeled datasets used to train the SOD classifier models. Method indicates the SOD scoring method used to label the different anatomical regions. NoL gives the number of labeled images for the particular dataset.	53
4.3	Landis and Koch's [26] Fleiss' Kappa interpretation criteria.	56
4.4	Megyesi et al.'s [33] SOD classification results.	57

4.5	Gelderman et al.'s [17] SOD classification results.	58
4.6	The interrater study results. Reported are the Fleiss' kappa value, the p-value, the upper bound (UB) and lower bound (LB) of the 95% confidence interval (CI), and the agreement level for both the human-human and AI-human agreement across both scoring methods. The agreement levels are determined using Landis and Koch's [26] interpretation criteria, as shown in Table 4.3.	59
5.1	The univariate and multivariate linear regression analysis results. Reported are the adj. R^2 , the RMSE, and the back-transformed (i.e., from the log scale back to the normal scale) formula with the standardized regression coefficients.	80
B.1	Megyesi et al.'s [33] stages and description for the head (including the neck).	103
B.2	Megyesi et al.'s [33] stages and description for the torso.	104
B.3	Megyesi et al.'s [33] stages and description for the limbs.	105
B.4	Gelderman et al.'s [17] scoring method for the head. Point.1, point.2, etc. represent the different phenomena.	106
B.5	Gelderman et al.'s [17] scoring method for the torso. Point.1, point.2, etc. represent the different phenomena.	107
B.6	Gelderman et al.'s [17] scoring method for the limbs. Point.1, point.2, etc. represent the different phenomena.	108

List of Figures

2.1	An overview of the Waterfall-Agile software development life cycles approach used for the development of Image Cloud Platform for Use in Tagging and Research on Decomposition. In the figure, the black boxes and arrows represent the linear phases of the Waterfall method, while the blue arrows illustrate the integration of the Agile method to enable the revisiting of phases as needed and to allow multiple phases to be worked on simultaneously, making it easier to respond to changing product requirements.	13
2.2	An architecture diagram illustrating the construction and interplay of the various components of the Image Cloud Platform for Use in Tagging and Research on Decomposition web application. When a user interacts with the web browser user interface on the client-side, a request is generated. This request is first processed by AngularJS before being sent to the server-side, where it is managed by NodeJS. Subsequently, it is directed to Express.js, which communicates with the MongoDB database to retrieve the requested data. MongoDB retrieves the data and sends it back to Express.js, which forwards it to NodeJS, and in turn, to AngularJS, which finally displays the data to the user through the web browser.	16

2.3	The Image Cloud Platform for Use in Tagging and Research on Decomposition user interface showing the search-by-subject feature. Each thumbnail represents one donor along with the anonymized donor ID (e.g., donor 88a: first thumbnail in the first row), the total number of days photos have been taken for that donor, and the total number of images of that donor. Images of specific donors can be searched for using the “Search by ID” search bar. The “Sort by” dropdown menu can be used to sort the donors either by (1) the total number of days or (2) the total number of photos.	22
2.4	The Image Cloud Platform for Use in Tagging and Research on Decomposition user interface showing the search-by-subject drill-down feature, which allows users to browse images of a specific donor. For example, when the user clicks the donor thumbnail with donor ID 005 (left red box), the user is presented with the “Days” view (follow “Arrow 1”) where each thumbnail represents a specific day along with the date (e.g., date 00808 enclosed by the right red box, which follows the pattern of [x][mmdd], where x is the number of years in the ARF, mm the month, and dd the day) and the total number of photographs taken on that date. From here, a user can search for a specific day by using the “Search by date” search bar or use the “Sort by” dropdown menu to sort the day thumbnails by (1) the date in ascending order, (2) the date in descending order, or (3) the total number of photos in descending order. In addition, a user can click on a specific day, such as date 00808 (right red box) to be presented with the “Images” view (follow “Arrow 2”), which displays all photographs for that day. From this point, a user can click on an image and be presented with a larger version, which can directly be tagged as demonstrated by Figure 2.6.	23

2.5	The Image Cloud Platform for Use in Tagging and Research on Decomposition user interface showing the search-by-image-content feature. Here, a user can search for previously tagged images by either using the “Text Search” or “Checkbox Search” option. The “Text Search” allows a user to enter multiple tags separated by a comma without any space (e.g., tag1,tag2,tag3), and select the option between “AND” (i.e., “logical AND” operator) and “OR” (i.e., “logical OR” operator), while the “Checkbox Search” allows a user to select tags from the list of pre-existing tags and select the option between “AND” (i.e., “logical AND” operator) and “OR” (i.e., “logical OR” operator). Once one of the search buttons is hit, all images that were previously tagged with the chosen tags are presented to the user.	24
2.6	Demonstration of Image Cloud Platform for Use in Tagging and Research on Decomposition’s tagging feature. Here, a user drew a polygon around the area was then presented with a text box where he/she entered “fly”, selected the tag, and hit the “Save” button to create the tag. The “Cancel” button can be used to delete the current polygon and restart. Note: If the keyword that was entered into the text box does not exist in the predefined nomenclature, no tag can be created for that keyword.	25

2.7	The pre-study survey results summarized as a stacked area graph. Specifically, the graph shows each participant's (x-axis; User1 = participant 1, User2 = participant 2, etc.) total familiarity score (y-axis), ranging from 1 (not familiar) to 20 (very familiar). The integers within each colored region indicate a user's familiarity score for the different concepts. For example, "User1" has the familiarity scores 4, 4, 2, and 2 for the concepts human decomposition, forensic work, image db, and multimedia curation, receptively, resulting in a total familiarity score of 12. It can be seen that the lowest familiarity score was 9 and the highest was 17, with most of the users being more familiar with human decomposition and forensic work, and less with multimedia curation and image databases.	27
2.8	The post-study survey results summarized as a stacked area graph. Specifically, the graph shows each participant's (x-axis; User1 = participant 1, User2 = participant 2, etc.) total difficulty score (y-axis), ranging from 1 (very difficult) to 15 (very easy). The integers within each colored region indicate a user's difficulty score for the different tasks. For example, "User1" has the task difficulty scores 5, 5, and 5 for the tasks 1–3, respectively, resulting in a total difficulty score of 15. It can be seen that the lowest total difficulty score was 5 and the highest total difficulty score was 15, with most of the participants rating Tasks 1 and 2 as the easiest.	28
3.1	Overview of the developed SOD-VS with the proposed label propagation method.	35
3.2	No decay. The image has been blurred to make them less disturbing to sensitive viewers.	37
3.3	Fully decayed. The image has been blurred to make them less disturbing to sensitive viewers.	38

4.1	The data processing pipeline.	51
4.2	The SOD classification framework was trained using a two-step transfer learning approach. In step 1, the pre-trained convolutional layers of the base model (blue rectangle) were frozen, and only the newly added classifier layers (yellow rectangle) were trained. In step 2, the previously frozen layers from step 1 (blue rectangle) were unfrozen, and the entire model was trained end-to-end.	54
5.1	Study sample (n=213) distribution plots by (a) sex, (b) age, (c) BMI group, and (d) PMI.	69
5.2	Plot of TDS vs. PMI (n=213).	74
5.3	Plot of TDS vs. naturally log-transformed PMI (n=213).	75
5.4	Plot of TDS vs. ADD (n=213).	76
5.5	Plot of TDS vs. naturally log-transformed ADD (n=213).	77
5.6	Predicted vs. actual PMI scatter plots for the different regression analyses with input variables: (a) TDS, (b) TDS + demographic features, (c) TDS + demographic features + weather features.	82
5.7	Predicted vs. actual ADD scatter plots for the different regression analyses with input variables: (a) TDS, (b) TDS + demographic features, (c) TDS + demographic features + weather features.	84
A.1	The complete nomenclature consisting of 132 forensic-related keywords.	101

A.2	The ETL data pipeline illustrates ICPUTRD’s data management process. The first step was to extract the raw data, which included all donor images and image metadata, such as donor demographics and temperature and humidity recordings from Tinytag data loggers. Once the raw data was extracted, it was processed, cleaned, and transformed to fit the required structure. Specifically, this involved standardizing filenames, anonymizing donor IDs (i.e., assigning a unique 3-digit hexadecimal number), removing duplicate images using image hashing [20], correcting image orientation, and detecting and removing images containing irrelevant content (e.g., stakes displaying unanonymized donor IDs or plastic bags/cages covering the donor) using a trained CNN classifier to automate this task [36]. Finally, the transformed data was loaded into the target system, involving local storage as well as storage in a MongoDB database.	102
-----	---	-----

Chapter 1

Introduction

1.1 Problem Statement

Determining the post-mortem interval (PMI), which is the time between death and when the body was found, is an important forensic anthropology problem. It involves estimating the time human remains have been decomposing based on the inspection of a cadaver. An accurate estimation of the PMI can significantly narrow down the list of potential decedents the remains might correspond to, facilitating the eventual identification of the individual. In cases involving homicide, law enforcement professionals can utilize the PMI to eliminate potential suspects and validate witness accounts. Additionally, having knowledge of the PMI aids in establishing the spectrum of natural occurrences and environmental influences that impacted the remains as the seasons passed, thereby enabling a more comprehensive analysis.

The Forensic Anthropology Center (FAC) at the University of Tennessee, Knoxville (UTK), is home to the Anthropology Research Facility (ARF), where human decomposition has been studied since 1981. Since 2012, multiple photographs of each donor (approximately 100 donors annually) have been taken daily throughout the entire decomposition process. To date, the database contains over one million photographs obtained from more than 800 donors. Since 2016, the Department of

Electrical Engineering and Computer Science and the Department of Anthropology at UTK have been collaborating on a project funded by the National Institute of Justice (NIJ) with two key goals: (1) to make this large human decomposition photographic collection usable for forensic research and casework and (2) to develop artificial intelligence (AI) tools for forensic purposes, with the big goal being PMI estimation.

Making this photographic collection accessible has the potential to transform significant aspects of forensic work. It enables the reuse of existing digital data in forensic investigations or research, allowing for more extensive and accurate studies at the same cost as reproducing conditions in a physical experiment. Additionally, it provides the building blocks for downstream tasks, such as developing AI techniques for forensic purposes.

A central part of this research is to contribute to those efforts and advance the research of using AI for forensic purposes. In particular, the contributions of my dissertation are summarized as follows:

1. Continue to develop a collaborative tool for forensic research and casework that facilitates the organization, visualization, analysis, and curation of a large collection of human decomposition photographs.
2. Develop a method to perform stage of decay labeling when labeled data is limited, and subsequently, develop an AI method to automate stage of decay identification.
3. Evaluate and improve upon existing PMI estimation methods using our human decomposition data set, which allows for a much larger sample size than earlier studies.

1.2 Background and Literature Review

Human decomposition is a natural and continuous process involving the breakdown of tissues after death that can span from weeks to years, contingent upon a range of biological and environmental factors. Comprehending how these factors influence the rate of human decomposition is pivotal in ascertaining the PMI, which is a crucial piece of information in forensic contexts and something that has long been sought after. An accurate estimation of the PMI can streamline the pool of missing persons data to be investigated and can also establish a precise timeline for the entire case [57, 33, 17, 16].

Since there is no direct measure for human decay, researchers have developed methods that categorize the human decomposition process into stages. Each stage is characterized by distinct decomposition phenomena, aiding in the estimation of the PMI [50, 16]. Megyesi et al. [33] employed a classification system that separates human decomposition into three distinct anatomical regions (i.e., head and neck, torso, and limbs), which were further subdivided into four high-level sequential stages: fresh, early decomposition, advanced decomposition, and skeletonization. Within each stage, specific decomposition phenomena are enumerated and assigned point values if present. The scores of each of the three anatomical regions are utilized in a calculation devised in the study to estimate the PMI. Building upon the work of Megyesi et al. [33], Gelderman et al. [17] categorizes each anatomical region into six sequential stages, with the lowest indicating no visible changes and the highest indicating complete skeletonization. Similar to Megyesi et al. [33], these regional scores are utilized to estimate the PMI.

Presently, establishing a corpse’s stage of decay is primarily done manually by trained experts using established evaluation methods, such as those proposed by Megyesi et al. [33] and Gelderman et al. [17]. Non-metric methods, such as determining the stage of decay, which rely on subjective judgments made by human experts, might potentially introduce bias and errors [38, 10, 25]. Over the years,

artificial intelligence techniques, particularly machine learning, have been employed to assist human decision-making. AI assesses information based on the entirety of acquired facts or data using advanced algorithms to consistently provide accurate and objective insights, thereby reducing the likelihood of human error and bias [47]. To the best of our knowledge, the automation of human decay staging methods using AI techniques has not been previously attempted. This may serve as a proof-of-concept for future studies on the advancement and integration of AI in forensic practice and research, particularly for the estimation of the PMI.

Once the stage of decay has been identified, experts use this piece of information to help them estimate the PMI. However, as mentioned earlier, decomposition varies due to many factors. Temperature is believed to be one of the most influential "independent" variables in the human decomposition process [28]. The use of accumulated degree days (ADD), routinely employed in the practice of forensic entomology to predict insect development, has provided a solution to model the effects of temperature on decomposition [37]. Specifically, instead of solely predicting the PMI based on observed levels of accumulated decomposition, which considers only time, the ADD is estimated, incorporating both time and temperature, which is then used to back-calculate the PMI. Megyesi et al. [33] and Gelderman et al. [17] both applied this approach in their studies to develop formulas for PMI and ADD estimation. The sample sizes used to develop these formulas by both Megyesi et al. [33] (57 outdoor cases) and Gelderman et al. [17] (12 outdoor cases) are small compared to our human decomposition data set, which covers over 800 outdoor cases. Evaluating a method on a larger sample will better represent the population and, hence, increase the statistical power of the evaluation. This, in turn, can provide novel insights that may further research in developing methods for PMI estimation, particularly from human remains.

1.3 Research Objectives

The proposed dissertation will focus on the application and development of methods to address the challenges and requirements described in Sections 1.1 and 1.2. Specifically, this dissertation will include four chapters centered around the following research questions:

1. Can we develop and evaluate a collaborative tool for data organization, analysis, and annotation to facilitate not only this research study but also others in forensic research and casework? This tool, in particular, should enable researchers to easily access, enhance (or curate), and analyze the extensive human decomposition data set. All design and implementation steps taken to develop the tool will be described. Additionally, to evaluate the tool's suitability for all its intended users and use cases, a user study will be conducted and reported. The final tool will facilitate the work of the remaining dissertation chapters.
2. Can we train a model on the human decomposition image data set to predict the stage of human decay? Human decomposition is typically divided into broader categories or linear stages (e.g., fresh, early decomposition, advanced decomposition, and skeletonization), each characterized by different decomposition phenomena. In order to train a model to learn the different stages of human decay, a labeled data set is required, which in turn requires domain expertise. Because of a shortage of human resources with expertise in this domain, there is a limitation that presents a challenge in terms of the availability of labeled data for model development. In this chapter we tackle learning with limited labeled data.
3. This chapter extends the research from question 2, where the goal is to explore the feasibility of automating two established human decomposition scoring methods using AI. Additionally, an interrater study will be conducted to assess

the reliability of the AI models compared to human forensic examiners for stage of decay identification. This chapter seeks to answer the following questions: (1) Which human decay staging method is more effective for training the AI model?, (2) How do the trained AI models compare to different human raters?, and (3) What is the level of reliability among human raters when applying the different decomposition scoring methods?

4. Once the stage of decay has been identified, experts use this information to help estimate the PMI. However, temperature has been identified as one of the most important factors in the human decay process. Therefore, researchers commonly estimate PMI proxies, such as accumulated degree days (ADD), which take into account both time and temperature, potentially leading to a more accurate PMI prediction. Existing decomposition-based PMI/ADD estimation methods were either developed or evaluated on a rather small sample size. Additionally, they did not consider factors such as donor demographics and environmental information, which are known from existing literature to affect the human decay process. Therefore, this chapter aims to investigate ways to improve decomposition-based PMI prediction by addressing two key questions: (1) How do existing PMI/ADD estimation methods perform when applied to the significantly larger sample size available to us? and (2) Can we improve upon these methods by incorporating additional data known to affect human decomposition?

1.4 Dissertation Structure

The dissertation follows a four-paper structure. That is, the four core chapters (2,3,4,5) have either been published, submitted to a journal for review, or are in the process of being submitted to a journal. The final chapter (6) summarizes the

main contributions and findings of this work and includes recommendations for future research.

Chapter 2

A Collaborative Tool for Human Decomposition Research and Casework

Disclosure Statement

A version of this chapter was originally published in the Journal of Forensic Sciences:

A.-M. Nau, S. Mousavi, D. Lee, R. Hossain, T. Griffin, D. W. Steadman, and A. Mockus. Icputrdr: Image Cloud Platform for use in tagging and research on decomposition. Journal of forensic sciences, 2023.

My contributions: Continuous software development, including integration, deployment, delivery, and testing, to adapt to the needs of the project and users. ETL data pipeline development and maintenance, which involved data extraction, data cleaning and processing, and data loading (see Appendix [A.2](#) for details). Responsible for publication efforts.

2.1 Abstract

Human decomposition studies aim to understand the various factors influencing human decay to assess the deceased and develop post-mortem interval (PMI) estimation methods. These types of studies are typically conducted through physical experiments examining the deceased; however, big data systems have the potential to transform how large-scale forensic anthropology research questions can be addressed with curated images of donors with known demographic, climatic, and post-mortem historical data. This study introduces ICPUTRD (Image Cloud Platform for Use in Tagging and Research on Decomposition), a web-based software system, which enables forensic scientists to easily access, enhance (or curate), and analyze very large photographic collections documenting the longitudinal process of human decomposition. ICPUTRD, a JavaScript-based application, was designed and built through a combination of the Waterfall and Agile software development life-cycle methods and provides an image search and tagging features with a predefined nomenclature of forensic-related keywords. To evaluate the system, a user study was conducted, involving 27 participants who completed pre- and post-study surveys

and three research tasks. Analysis of the study results confirmed the feasibility and practicality of ICPUTRD to facilitate aspects of forensic research and casework involving large collections of digital photographs of human decomposition. It was observed that the nomenclature lacked certain law enforcement keywords, so future work will focus on expanding it to ensure ICPUTRD is suited for all its intended users.

2.2 Introduction

Human decomposition studies aim to identify both the individual and environmental factors affecting human decay to be able to assess unique features of the deceased as well as develop methods for post-mortem interval estimation (PMI), an important aspect of forensic investigation[58, 7]. The Forensic Anthropology Center (FAC) at the University of Tennessee, Knoxville (UTK), houses the Anthropology Research Facility (ARF), popularly known as the “Body Farm,” which is the first outdoor decomposition laboratory. Individuals can choose to register for self-donation of their bodies during their lifetime, while legal next of kin can also choose to donate the bodies of their loved ones through the FAC Body Donation Program. Each year, approximately 100 donors are placed at the ARF to facilitate investigations of human decay. Since 2011, photographs of these donors have been taken daily from the moment they are placed at the facility until advanced decay and skeletonization occur, whereby photos were taken weekly or monthly. Until recently, this large photographic collection was stored in individual file folders, making analytical tasks difficult and tedious. However, if this valuable image data were more accessible, it has the potential to transform fieldwork-intensive research. For instance, unlike physical donors, images can be re-examined at any given time and from various perspectives depending on the research study, and at a potentially lower cost when compared to physical experiments. In addition, such a large photographic collection offers the opportunity to extrapolate from larger and statistically relevant samples, thereby

supporting rigorous quantification of critical forensic estimates, such as the PMI—a significant challenge in forensic investigation [33, 29].

Making large image collections in the forensic decomposition domain suitable for research and casework has received little attention. This paper aims to introduce the Image Cloud Platform for Use in Tagging and Research on Decomposition, abbreviated as ICPUTRD, a big data, collaborative, cloud-based image platform developed to address this limitation and thereby facilitate research on human decomposition. The main objective of this platform is to enable the analysis into how both intrinsic factors (e.g., age, biological sex, body composition, etc.) and extrinsic factors (e.g., temperature and moisture, animal scavenging, clothing, etc.) impact the decomposition process in a consistent and replicable manner. Specifically, ICPUTRD was designed for two key purposes: (1) to enable image searches using both pre-existing and newly generated metadata, including demographic information (i.e., age, sex, body weight, etc.), temperature and humidity data, and forensic-related keywords (i.e., anatomical, stages of decomposition, presence of scavenging, etc.), and (2) to enhance existing images through an image tagging feature that simultaneously enriches the metadata associated with the images. Notably, the platform’s tagging feature allows individual images to be labeled (or tagged) with relevant decomposition information using a predefined set of forensic-related keywords, thereby generating new image metadata. ICPUTRD was developed using a combination of JavaScript-based technologies and includes a user-friendly interface. The source code and documentation can be found at <https://bitbucket.org/icputrd/platform>.

To the best of our knowledge, this is the first tool that met all system requirements. Specifically, the tool should be capable of addressing the following three key tasks: (1) supporting large-scale data visualization, storage, and analysis, (2) facilitating data curation (i.e., organizing, enhancing, and preserving data), and (3) offering extensibility (i.e., allowing for new functionalities or modifications of existing ones). Additionally, none of the available tools were designed and evaluated with the intention of being used for human decay-related use cases. Consequently, they may

be limited in their applicability, potentially requiring additional technical setups and further development.

2.3 Materials and Methods

ICPUTRD was designed and built through a collaboration of two software development life cycles (SDLC), Waterfall and Agile methods, as shown in Figure 2.1. Specifically, the Waterfall methodology was used to make a high-level design plan, which involved the following four sequential phases: (1) requirement gathering, (2) design, (3) implementation, and (4) system evaluation. Unlike Waterfall, which follows a linear process (i.e., a phase cannot commence until the previous phase has been completed, nor can it be revisited once completed), the Agile method is a more iterative approach that does not demand one phase to conclude before the next one begins, but allows for multiple phases to be worked on simultaneously and permits revisiting phases later in the development life cycle to respond to changing user requirements [19, 3, 8]. Details on the requirement gathering, design, implementation, and system evaluation phases are provided in the following sections.

2.3.1 Requirement Gathering

The requirements of ICPUTRD were defined based on its use cases, including the ability to search for and analyze (1) images of a specific donor and date and (2) images containing specific forensic-related features (i.e., anatomical, stages of decomposition, presence of scavenging, etc.). In order for ICPUTRD to be a system that can facilitate such use cases, it must meet the following two main requirements: (1) an image search feature, providing the capability to search for and browse images using different search filters, and (2) an image tagging feature, allowing users to further enrich the data by annotating images with forensic-relevant information. The first requirement concerns effective searching essential to find relevant photos among over one million images.

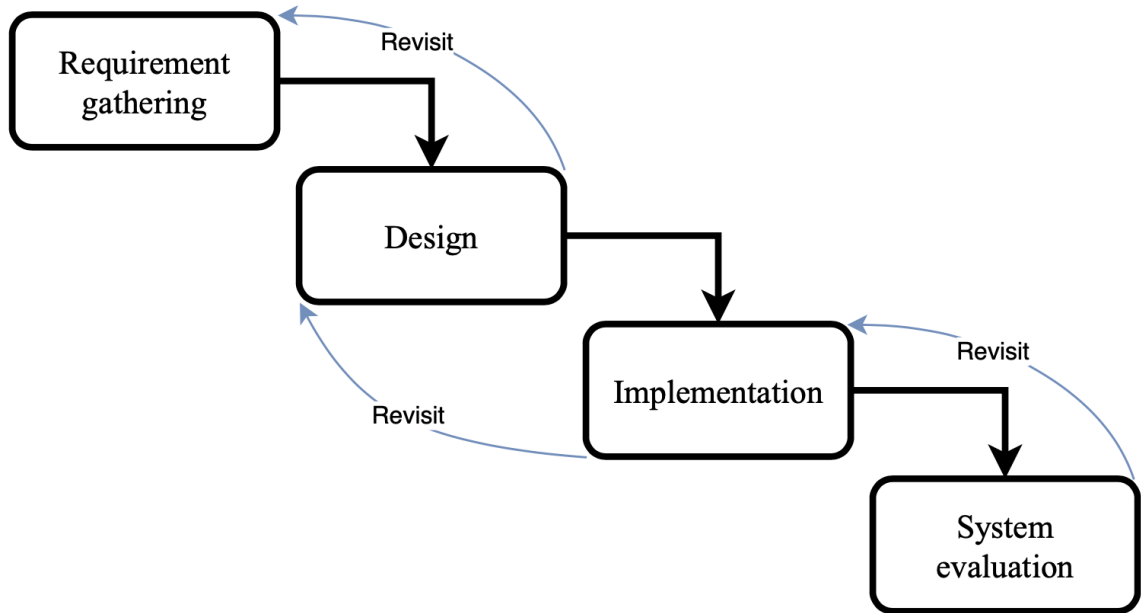


Figure 2.1: An overview of the Waterfall-Agile software development life cycles approach used for the development of Image Cloud Platform for Use in Tagging and Research on Decomposition. In the figure, the black boxes and arrows represent the linear phases of the Waterfall method, while the blue arrows illustrate the integration of the Agile method to enable the revisiting of phases as needed and to allow multiple phases to be worked on simultaneously, making it easier to respond to changing product requirements.

Specifically, the system’s search feature should provide the ability to find images based on the subjects, demographic information, and temperature and humidity data associated with the time of decomposition, and the content of the images themselves (e.g., images of heads or images containing flies, mummified tissue, etc.). Since the metadata associated with the human decomposition images only includes the donor ID and the time a particular image was taken, the image data needs to be enhanced for a user to be able to make content-based image searches, which is part of the first requirement. The second system requirement aims to address this limitation through a process called tagging, which is where a particular area of an image is “tagged” or “labeled” with a keyword. More specifically, the system should provide users with a suitable interface that facilitates this enrichment or tagging process via a nomenclature. This nomenclature of keywords should be carefully defined by domain experts to ensure that the metadata created during tagging is both useful and relevant for forensic research and casework, especially in cases where PMI is part of the analysis. In addition, a solid and predefined set of keywords ensures consistency across different users and prevents misspellings during tagging. Finally, the created image tags should be stored reliably and available for viewing, editing and deletion when needed.

2.3.2 Design and Implementation

Once the requirements were gathered, a high-level design plan including the system architecture (i.e., the main system components and their relationships) and technology stack (i.e., programming languages, frameworks, data storage, front-end & back-end tools, etc.) were created with respect to the system’s two requirements: (1) image search feature and (2) image tagging feature for data enhancement. To be able to search for images, the information must be available; users should be provided with an interface that allows image searches by the nomenclature, and the interface should be able to communicate with the stored data, extract the necessary

data, and present it to the users. As a result, ICPUTRD’s search feature consists of three main components: (1) the client-side: a user interface that enables user-friendly image searching, (2) the database system: a program that facilitates data storage, retrieval and modification at any time, and (3) the server-side: also called the back-end, which handles all incoming requests from the client-side and provides the logic for the communication between the client-side and the database. These components were implemented using the MEAN stack [32], a JavaScript-based framework for web application development, which is named after and powered by MongoDB [35] as the database server, ExpressJS [14] as the server-side JavaScript framework, AngularJS [1] as the client-side JavaScript framework, and NodeJS [44] as the server-side JavaScript runtime.

The user interface resides on the client-side within a web browser and was implemented using AngularJS. It can be accessed across the web by permitted users to search for images. ICPUTRD uses MongoDB, an open-source, cross-platform database, as its storage system. It stores data in key-value pairs, using the widely used JSON format. Specifically, MongoDB was used to store all human decomposition image paths along with its metadata, and all users who have access to ICPUTRD along with their user activity. ICPUTRD’s server-side, which oversees all logical operations and computations in response to user interactions, was implemented using NodeJS and Express.js, two back-end JavaScript web application frameworks. The architecture diagram shown in Figure 2.2 provides an overview of how these different software components interact. When users interact with the web browser user interface, their request is initially processed by AngularJS on the client-side. Subsequently, it is routed to the server-side, where NodeJS manages it before forwarding it to Express.js. Express.js then communicates with the MongoDB database to retrieve the requested data. MongoDB fetches the data and sends it back to Express.js, which, in turn, forwards it to NodeJS and then to AngularJS where it is then displayed to the user through the web browser.

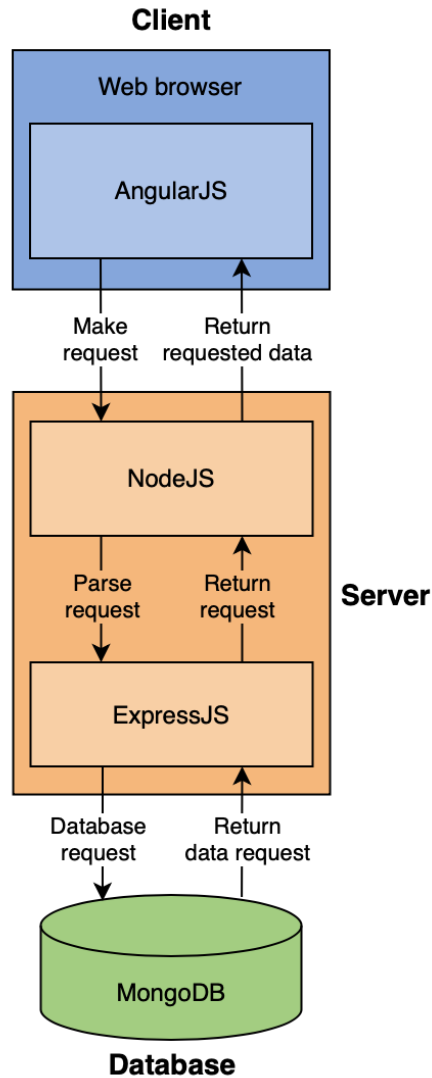


Figure 2.2: An architecture diagram illustrating the construction and interplay of the various components of the Image Cloud Platform for Use in Tagging and Research on Decomposition web application. When a user interacts with the web browser user interface on the client-side, a request is generated. This request is first processed by AngularJS before being sent to the server-side, where it is managed by NodeJS. Subsequently, it is directed to Express.js, which communicates with the MongoDB database to retrieve the requested data. MongoDB retrieves the data and sends it back to Express.js, which forwards it to NodeJS, and in turn, to AngularJS, which finally displays the data to the user through the web browser.

ICPUTRD’s tagging feature was also implemented using the MEAN stack, but in addition, the Annotorious library [2], a JavaScript image annotation library, was used. Like the platform’s search feature, the tagging feature involves the client-side, server-side, and database. Specifically, the tagging interface (i.e., the client-side) displays an image to the user in full size through the web user interface. When the interface loads, the server-side accesses the MongoDB database to look for any existing tags associated with the displayed image to show to the user. The user can then view and modify any existing tags as well as create new tags for the displayed image. As mentioned before, the tags created come from a predefined nomenclature of keywords, which was developed by forensic experts from the ARF ensuring the keywords were relevant for various types of end users and research studies. Specifically, the development of the nomenclature involved an iterative process of forensic experts tagging a set of images with forensic information, including anatomical (e.g., head, torso, leg, etc.), decomposition (e.g., bloat, fresh, mummification, etc.), insect activity (e.g., fly, insect, maggot, etc.), and scavenging information. This was followed by the computer and data scientists extracting the tags, sorting by frequency (in descending order), and detecting and normalizing synonyms (e.g., “mummification” was chosen instead of “mummified”, or “mummified skin”). The final nomenclature of keywords was then implemented into ICPUTRD to accommodate its tagging and search features.

2.3.3 System Evaluation

To ensure that the implementation of ICPUTRD satisfies all the defined requirements and to investigate whether the target audience of forensic scientists finds the system acceptable for conducting several research tasks, an IRB approved user study was designed and executed. The user study involved four main steps: (1) study design, (2) recruitment of participants, (3) execution of the study, and (4) analysis of the results.

An empirical evaluation method, following Nielson [43], was applied to assess three system usability measures: (1) effectiveness (i.e., a user’s ability to complete the tasks), (2) efficiency (i.e., the number of resources used to complete the user study tasks), and (3) satisfaction (i.e., a user’s experience with interacting with the system) [4, 5]. More specifically, potential users were exposed to the system by completing three research tasks designed with the goal of evaluating ICPUTRD’s search and tagging features. The first task attempted to gauge the effectiveness of the current implementation with a simple task of finding images by specific tags. The purpose of the second task was to evaluate the range of keywords potential users may be interested in. The aim of the third task was to evaluate ICPUTRD’s tagging feature. It is important to note that prior to the user study, at least one tag for each keyword from the predefined nomenclature was created by forensic experts to ensure the image search tasks could be performed successfully. The exact user study task instructions were as follows:

1. “Search for images using keywords (up to five) from the predefined nomenclature.” The purpose of this task was to evaluate ICPUTRD’s search feature and user-friendliness.
2. “Using up to five keywords of your choice (different from the five from task 1), search for images that you might want to use for your research or work, or both.” The purpose of this task was to see if any keywords were missing from the current predefined nomenclature and to determine what keywords were most frequently used by the participants.
3. “Given the provided image, create five tags.” For this task all participants were shown the same image. The purpose of this task was to evaluate the usability of ICPUTRD’s tagging feature.

In addition to the above-described tasks, the user study involved a pre-study survey to get a general idea about the background of the users on certain concepts,

and a post-study survey to learn about the users’ experience from interacting with the system and performing the tasks. The pre-study survey collected information on a participant’s level of familiarity with the following four concepts: multimedia curation, image database (db), forensic work, and human decomposition. Specifically, each concept was given a score ranging from 1 (not familiar) to 5 (very familiar), which were then added to obtain a participant’s final familiarity score. The post-study survey collected information on (1) whether a participant was able to complete each task, and (2) the difficulty they associated with each task ranging from 1 (very difficult) to 5 (very easy), which were again added to obtain each participant’s total difficulty score.

The participants of the study were recruited from the target population of ICPU-TRD users, including forensic anthropology researchers, professors and students, and law enforcement officers. Each participant was given 30 minutes to complete the user study and surveys (pre- and post-study). This 30-minute timeframe was based on how long, approximately, it should take each participant to complete the user study tasks and surveys.

2.4 Results and Discussion

2.4.1 Requirement Gathering, Design, and Implementation

Over a period of approximately one month, the image search feature and image tagging feature were defined based on iterative discussions (virtual and face-to-face) with domain experts from computer and data science and forensic anthropology.

The design and implementation of the system took approximately six months. This involved the developers writing all the code of the system, testing individual components for usability, and installing the system for its intended use. The forensic experts from the ARF provided the human decomposition image data along with its metadata. The images included overall images of donors, close-ups (e.g., head,

torso, arm, legs, etc.), and forensic items of interest (e.g., animal scavenging), and the metadata included donor ID and date, demographic information, and temperature and humidity data associated with each image. The amount of storage needed to store the compressed images was approximately 4TB. Storing images that were planned to be both edited or augmented would increase the storage needs approximately eight-fold, requiring at least 32 TB of storage, resulting in very slow and inefficient image search and retrieval.

Therefore, instead of storing all the images in the MongoDB database, they were stored directly on the file system and only their file paths were stored inside the database. Using a combination of MongoDB and the file system for image storage, drastically accelerated the data search and retrieval process. The carefully developed nomenclature of keywords consisted of 132 terms (see Appendix A.1). Table 2.1 shows example keywords of the final developed nomenclature. Note, the nomenclature is frequently updated through continuous user feedback loops.

The current version of ICPUTRD’s search feature can perform the following two types of searches:

1. Search-by-subject, as shown in Figure 2.3, allows users to search for all images associated with a particular subject or donor at a given time. A subject’s anonymized user ID (i.e., three-digit hex number) is used here as well as the system’s drill down feature (Figure 2.4) to find images for a specific date.

Table 2.1: Example keywords from the final nomenclature of forensic-related keywords.

Category	Keyword examples
Anatomic	head, torso, backside, leg, foot, hand, eye
Decay-related	fresh, bloat, early decomposition, advanced decomposition, skeletonization, mummification, purge
Other	fly, maggot, insect, scavenging, beetles, larvae

2. Search-by-image-content, as shown in Figure 2.5, allows users to search for images containing specific forensic-related keywords (e.g., hand, mummified, discolored skin, mold, maggots, liquid purge, etc.). More specifically, given a keyword, this search type retrieves all images previously tagged with that keyword using ICPUTRD’s tagging feature.

Note, since the image metadata also included demographic and temperature and humidity data, future work will focus on incorporating this data into ICPUTRD’s search feature. For now, this data is securely stored in the MongoDB database.

The ICPUTRD tagging feature is shown in Figure 2.6 where a “fly” tag was created. More specifically, the tagging feature works as follows: each click creates a vertex that will be connected to the next one until the user reaches the starting point and the polygon is closed. Once the polygon is completely drawn, a pop-up text box appears so that the user can type the desired keyword for the selected area. After typing the first three letters, all keywords from the predefined nomenclature that start with those letters are displayed to the user who then can either select the desired keyword, click the “Save” button to create the tag or click the “Cancel” button to discard the entire selection and start over. When a tag is created, the tag along with the polygon’s coordinates, the user’s ID, the date, and the file path of the image are stored in MongoDB for later retrieval using the search-by-image-content feature. The first version of the tagging feature only allowed rectangles to be drawn, but user feedback suggested that polygons would allow a more accurate tagging, and so changes were made accordingly.

2.4.2 System Evaluation via a User Study

A total of 27 participants were recruited for the user study, including one forensic anthropology researcher (MA degree), one professor from the Forensic Anthropology Department at UTK (PhD degree), one master and seven Ph.D. students from the

Donors

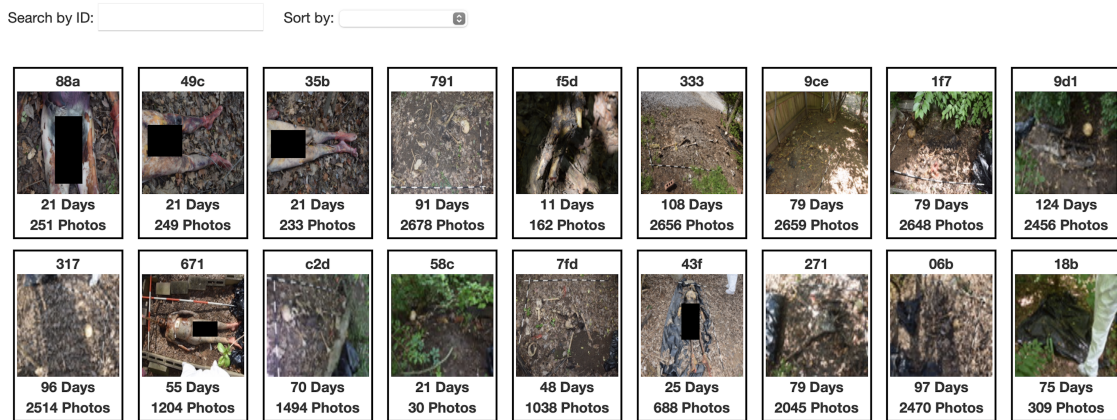


Figure 2.3: The Image Cloud Platform for Use in Tagging and Research on Decomposition user interface showing the search-by-subject feature. Each thumbnail represents one donor along with the anonymized donor ID (e.g., donor 88a: first thumbnail in the first row), the total number of days photos have been taken for that donor, and the total number of images of that donor. Images of specific donors can be searched for using the “Search by ID” search bar. The “Sort by” dropdown menu can be used to sort the donors either by (1) the total number of days or (2) the total number of photos.

Drill-down feature

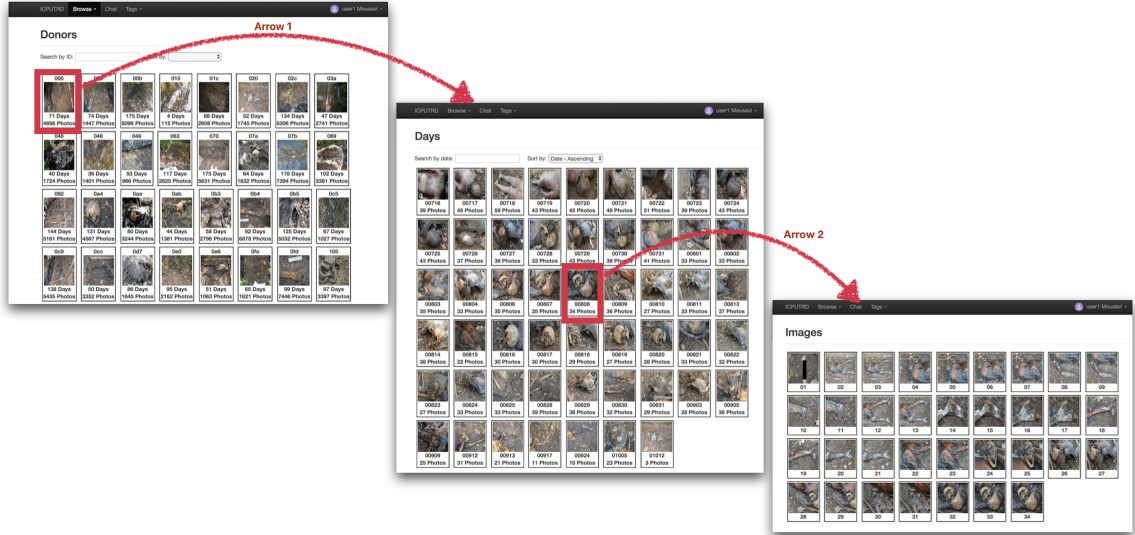


Figure 2.4: The Image Cloud Platform for Use in Tagging and Research on Decomposition user interface showing the search-by-subject drill-down feature, which allows users to browse images of a specific donor. For example, when the user clicks the donor thumbnail with donor ID 005 (left red box), the user is presented with the “Days” view (follow “Arrow 1”) where each thumbnail represents a specific day along with the date (e.g., date 00808 enclosed by the right red box, which follows the pattern of [x][mmdd], where x is the number of years in the ARF, mm the month, and dd the day) and the total number of photographs taken on that date. From here, a user can search for a specific day by using the “Search by date” search bar or use the “Sort by” dropdown menu to sort the day thumbnails by (1) the date in ascending order, (2) the date in descending order, or (3) the total number of photos in descending order. In addition, a user can click on a specific day, such as date 00808 (right red box) to be presented with the “Images” view (follow “Arrow 2”), which displays all photographs for that day. From this point, a user can click on an image and be presented with a larger version, which can directly be tagged as demonstrated by Figure 2.6.

Tag Search

Text Search

Enter comma seperated Tags, like- tag1,tag2,...

Select search type (AND will only show images with all searched tags present, OR will show images with one or more searched tags present):

AND

OR

Search

Clear

Checkbox Search

☐ fluid

☐ eggs

☐ scavenging

☐ marbling

☐ leg

☐ skin

☐ hand

☐ ear

☐ purge

☐ larvae

☐ skin slippage

☐ tattoo

☐ arm

☐ foot

☐ mouth

☐ mummification

☐ discoloration

☐ lividity

☐ mold

☐ bone

☐ eye

☐ genitals

☐ adult fly

☐ adipocere

☐ bloat

☐ insect

☐ head

☐ nose

☐ hair

Checkbox Search

Total has "0" images

Figure 2.5: The Image Cloud Platform for Use in Tagging and Research on Decomposition user interface showing the search-by-image-content feature. Here, a user can search for previously tagged images by either using the “Text Search” or “Checkbox Search” option. The “Text Search” allows a user to enter multiple tags separated by a comma without any space (e.g., tag1,tag2,tag3), and select the option between “AND” (i.e., “logical AND” operator) and “OR” (i.e., “logical OR” operator), while the “Checkbox Search” allows a user to select tags from the list of pre-existing tags and select the option between “AND” (i.e., “logical AND” operator) and “OR” (i.e., “logical OR” operator). Once one of the search buttons is hit, all images that were previously tagged with the chosen tags are presented to the user.

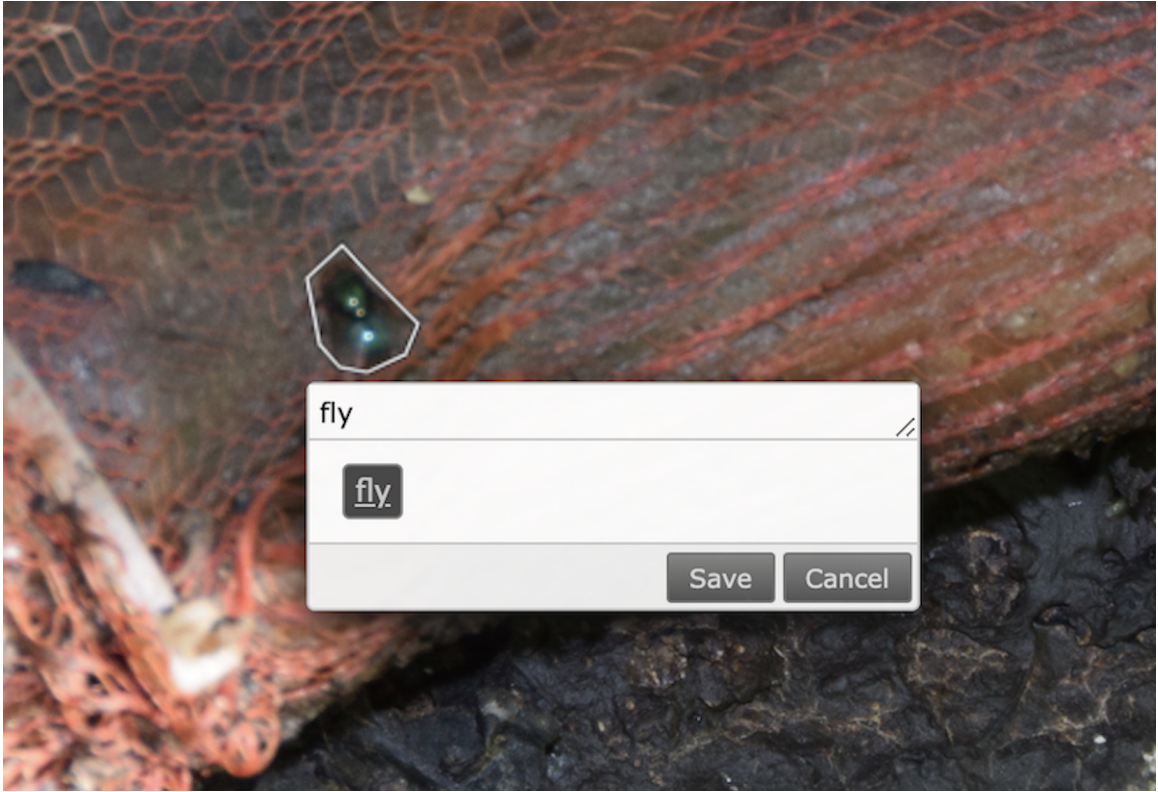


Figure 2.6: Demonstration of Image Cloud Platform for Use in Tagging and Research on Decomposition’s tagging feature. Here, a user drew a polygon around the area was then presented with a text box where he/she entered “fly”, selected the tag, and hit the “Save” button to create the tag. The “Cancel” button can be used to delete the current polygon and restart. Note: If the keyword that was entered into the text box does not exist in the predefined nomenclature, no tag can be created for that keyword.

Anthropology Department at UTK, 11 osteology students from the ARF, and six law enforcement officers.

The pre-study survey results are shown by Figure 2.7, which is a stacked area graph showing each participant's familiarity score, ranging from 1 (not familiar) to 5 (very familiar), and total familiarity score, ranging from 1 (not familiar) to 20 (very familiar). The lowest total familiarity score was 9 and the highest was 17, with most of the participants being more familiar with human decomposition and forensic work, and less with multimedia curation and image databases. Given that the educational backgrounds of the participants were mostly forensic- and law-related, this was expected.

The task difficulty scores recorded during the post-study survey are shown by the stacked area graph in Figure 2.8, ranging from 1 (very difficult) to 5 (very easy), and total difficulty score, ranging 1 (very difficult) to 15 (very easy). The lowest total difficulty score was 5 and the highest total difficulty score was 15, with most of the participants rating Task 1 and 2 as the easiest. Additionally, according to the post-study results, 27 (100%), 26 (93.2%), and 24 (88.9%) participants successfully completed Tasks 1, 2, and 3, respectively, given the 30-minute time limit.

Since Task 2 was not successfully completed by all participants, its results were further analyzed. Task 2 had two possible outcomes: (1) the participant was able to find at least one image for each of the five keywords of their choice, or (2) the participant was not able to find at least one image for each of the five keywords of their choice. Outcome (1) was recorded as success, while outcome (2) was recorded as failure and further analyzed, specifically the keyword(s) for which no images were found, meaning they were not included in the nomenclature of keywords. Through further analysis, keywords, such as "gunshot wounds", "knife wounds", "rigor mortis", "blister", "soldier fly", "pupae", and "foam," were discovered to not be in the existing nomenclature. It is important to note that majority of these keywords came from the law enforcement participants who did not participate in the development of

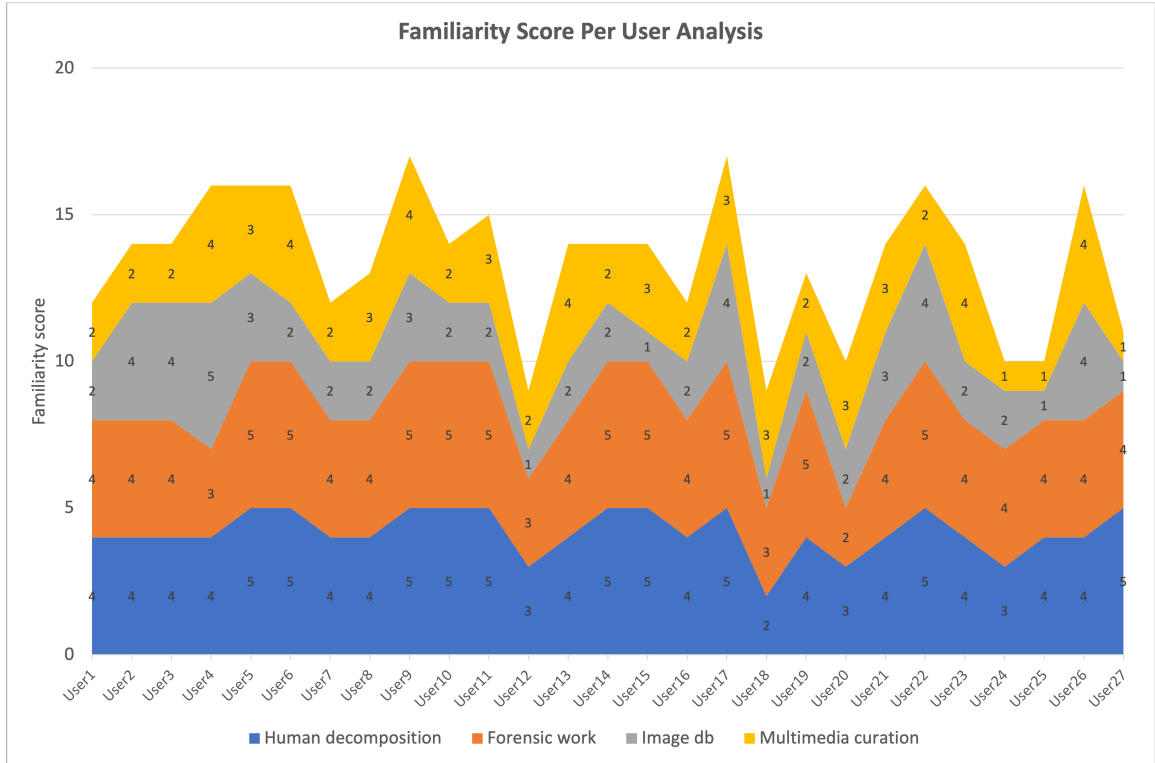


Figure 2.7: The pre-study survey results summarized as a stacked area graph. Specifically, the graph shows each participant’s (x-axis; User1 = participant 1, User2 = participant 2, etc.) total familiarity score (y-axis), ranging from 1 (not familiar) to 20 (very familiar). The integers within each colored region indicate a user’s familiarity score for the different concepts. For example, “User1” has the familiarity scores 4, 4, 2, and 2 for the concepts human decomposition, forensic work, image db, and multimedia curation, receptively, resulting in a total familiarity score of 12. It can be seen that the lowest familiarity score was 9 and the highest was 17, with most of the users being more familiar with human decomposition and forensic work, and less with multimedia curation and image databases.

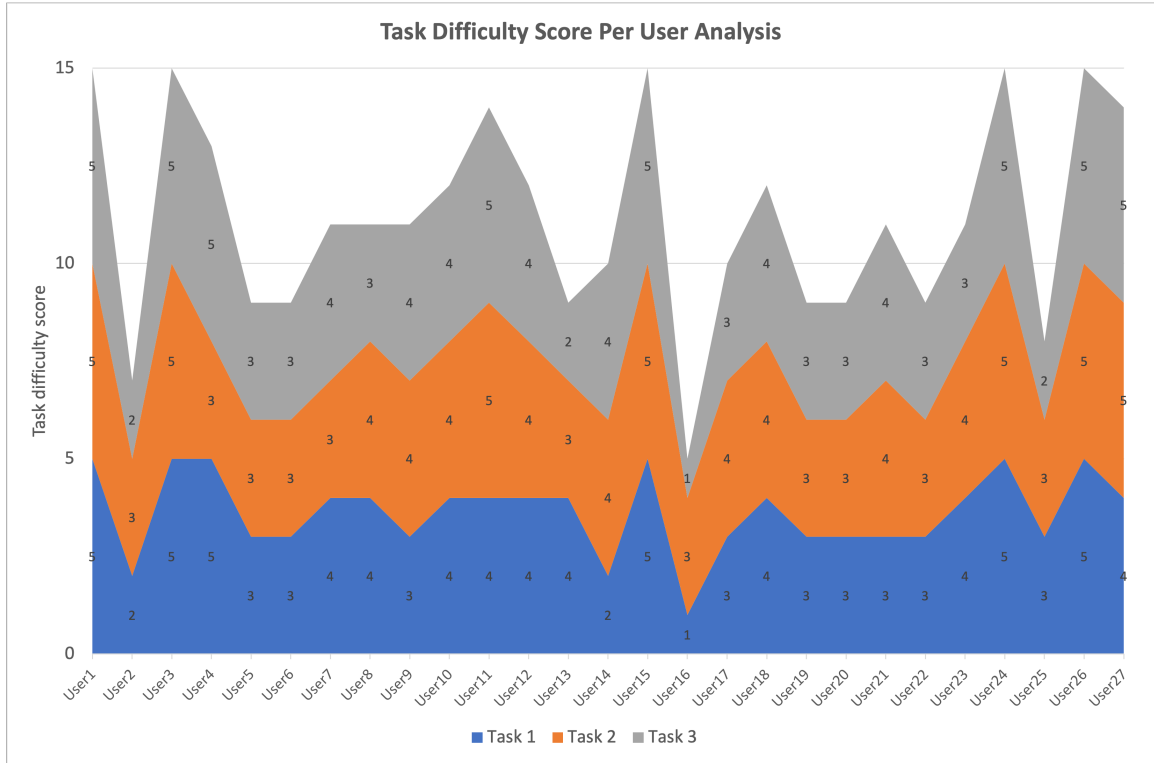


Figure 2.8: The post-study survey results summarized as a stacked area graph. Specifically, the graph shows each participant’s (x-axis; User1 = participant 1, User2 = participant 2, etc.) total difficulty score (y-axis), ranging from 1 (very difficult) to 15 (very easy). The integers within each colored region indicate a user’s difficulty score for the different tasks. For example, “User1” has the task difficulty scores 5, 5, and 5 for the tasks 1–3, respectively, resulting in a total difficulty score of 15. It can be seen that the lowest total difficulty score was 5 and the highest total difficulty score was 15, with most of the participants rating Tasks 1 and 2 as the easiest.

the nomenclature. These findings were used in future work to further improve the nomenclature to ensure ICPUTRD meets the needs of all target users.

Since the post-study survey showed that only 24 out of the total 27 participants were able to successfully complete Task 3, the results of the three unsuccessful participants were further analyzed. Based on this analysis, it was concluded that one of the following may be the reason for unsuccessful completion of Task 3: (1) a participant’s low level of familiarity with the concepts addressed in the pre-study survey (i.e., multimedia curation, image databases, forensic work, and/or human decomposition), (2) ICPUTRD is not sufficiently user-friendly, and/or (3) the 30-minute-timeframe was not long enough. Such findings will be considered in future work to further improve the usability of ICPUTRD’s tagging feature.

Furthermore, user study feedback suggested that manually performing certain administrative tasks, such as detecting privacy violations (e.g., images of stakes showing the donor ID), and removing irrelevant content (e.g., images of plastic bags covering the body), as well as domain tasks, such as body part and forensic feature identification, are limited from both a time and resource-perspective and suggest the need for process automation. The field of artificial intelligence (AI) provides a variety of automation technologies that can be used to analyze the human decomposition images, learn from them, and then perform certain manual tasks. Therefore, future work will focus on integrating AI technologies into ICPUTRD for automated data curation to reduce costs and increase efficiency.

In summary, the usability of ICPUTRD was found to be adequate for the tasks performed in the user study, and the keyword nomenclature was found to be sufficiently comprehensive to cover most basic scenarios. However, it was identified that the existing nomenclature lacked certain keywords, particularly terms related to law enforcement. To address this gap, further investigation will be conducted to determine the necessity of adding these missing keywords to the current nomenclature, ensuring its suitability for all intended users. Additionally, due to the time-consuming and costly nature of manual data curation, AI integration will be explored to automate

the process. It is worth noting that ICPUTRD is currently being deployed at the FAC and the Department of Anthropology at UTK, to contribute to both ongoing and new research efforts.

2.5 Conclusion

Historically, forensic research and casework have predominantly relied on physical experiments. The reported work aims to enhance the accessibility of forensic information through ICPUTRD, an image cloud-based platform that utilizes an extensive collection of digital photographs documenting the longitudinal process of human decomposition. The current implementation offers a search engine (specifically, search-by-subject and search-by-image-content functionalities) and a forensic-related image tagging tool, which enable easy data access, visualization, and enhancement. To assess ICPUTRD’s efficacy within the intended user community and purposes, a user study was conducted, which found it to be adequate for the study tasks and the keyword nomenclature to be sufficiently comprehensive for covering most basic use cases. However, since it was identified that the existing nomenclature lacked certain keywords, future work will focus on expanding the current nomenclature to ensure ICPUTRD is well-suited for all its intended users.

Chapter 3

Stage of Human Decay

Classification Given Limited

Labeled Data

Disclosure Statement

A version of this chapter was originally published in the proceedings of the 2023 IEEE International Conference on Image Processing (ICIP):

© 2023 IEEE. Reprinted, with permission, from A.-M. Nau, A. Mockus, and D. W. Steadman, Stage of Decay Estimation Exploiting Exogenous and Endogenous Image Attributes to Minimize Manual Labeling Efforts and Maximize Classification Performance, IEEE International Conference on Image Processing (ICIP), October 2023.

My contributions: Investigation, methodology, development, and writing. Specifically, I was in charge of data extraction and processing, model training and testing, and results analysis and reporting, including efforts related to publication.

3.1 Abstract

Establishing the stage of decay for human remains is an important and common task in forensic anthropology, which is presently done manually by trained experts. We aim to develop a supervised vision system, leveraging a large human decomposition image dataset, to perform stage of decay estimation. To minimize the manual labeling efforts and costs, we introduce and evaluate a label propagation method that exploits both exogenous and endogenous image attributes in conjunction with domain knowledge. The proposed label propagation method performed up to twice as fast (100%) as the baseline method and improved the classification macro-averaged F1 score of the best model from 0.79 to 0.863 (9.24%) when using three stage of decay classes and from 0.695 to 0.803 (15.54%) when using four stage of decay classes.

3.2 Introduction

Determining the stage of decay (SOD) is a vital and common task in forensic anthropology. Knowing the degree of decomposition can reduce the potential pool of

decadents that the human remains can belong to and help identify the deceased in human remain cases [17, 33]. Presently, establishing the SOD is done manually by trained experts via a thorough analysis of the corpse (e.g., soft tissue decomposition, skin discoloration, bone exposure, etc.), surrounding environment and climate, and presence of particular animals/insects. To the best of our knowledge, there is no research in the areas of computer vision for SOD estimation. We aim to make progress in this domain by attempting to build a supervised vision system to model human decay utilizing a large image collection collected over the past 10 years at the Forensic Anthropology Center at the University of Tennessee, Knoxville, which is home to the Anthropology Research Facility (ARF), also known as the “body farm”, an outdoor decomposition laboratory.

Manually labeling the entire image collection, particularly because domain expertise is required, is prohibitive from both a resource and time-perspective. As is common for domain-specific datasets, the ARF image collection contains external (or exogenous) image attributes, such as the donor ID and the date contained in the image filename. Furthermore, since the protocol requires photos of distinct body parts, it is known that the internal (or endogenous) image attributes, apart from the SOD, include particular body parts. While the exogenous attributes are explicitly associated with each image, the endogenous attributes are known only implicitly, that is, from the general protocol used to do the photography. As such, endogenous image attributes, in order to be used, need to be obtained or derived via some sort of image classification technique just as the SOD. Besides the exogenous and endogenous image attributes, we also have basic domain knowledge concerning human decomposition. The salient aspect of the domain knowledge is that the decomposition varies across subjects (or donors) and body parts and is gradual. In other words, for the photos taken in the same session (or date) of the same subject, the same body part should have exactly the same SOD. The key idea of our label propagation method was to leverage the exogenous and endogenous image attributes in conjunction with domain

knowledge to produce images similar to each image labeled by a forensic expert. The contribution of this work are summarized as follows:

1. A supervised vision system to perform SOD classification, utilizing a large human decomposition image collection, is presented and evaluated.
2. To reduce the need for costly and time-consuming manual labeling, a domain-aware label propagation method that leverages the exogenous and endogenous image attributes was developed. It is shown that our vision system using the proposed label propagation method provides improved SOD classification performance.

3.3 Proposed Vision System

This section presents the developed vision system (see Figure 3.1) we called SOD-VS for Stage of Decay Vision System. The key parts of SOD-VS are (1) data preparation, which involves manual data labeling, label propagation using the proposed method, and splitting data into train, validation, and test sets, and (2) a two-step transfer learning process to build and evaluate the SOD classification models. The SOD-VS implementation details and experimental settings are also presented.

3.3.1 The Human Decomposition Dataset

The dataset includes images of decomposing donors donated to the ARF. The images were taken by forensic experts at non-uniform intervals (one or more days apart) from various angles and of different body parts to show the different stages and areas of human decomposition. The dataset spans from 2011 to 2022, and includes over 1.5 million images from over 800 donors.

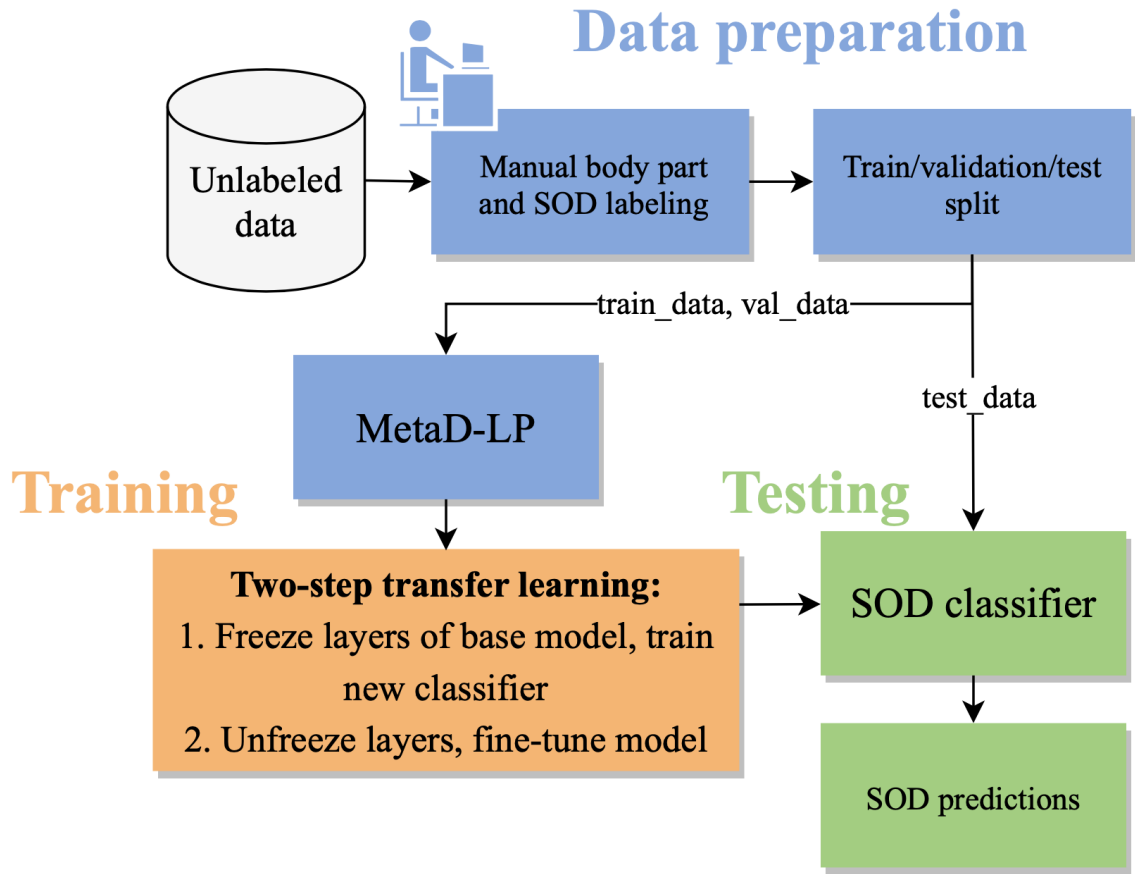


Figure 3.1: Overview of the developed SOD-VS with the proposed label propagation method.

3.3.2 Manual Data Annotation

The quality and size of the training data highly affects a model’s performance and generalizability. That is, the more representative and diverse the training data is, the more generalizable a model is likely to be. In the case of a temporal dataset, such as images documenting human decomposition in which the subjects’ appearance changes over time, it is important to sample the training data in such a way that it includes the dataset’s characteristics. In our case, for instance, images depicting all possible decomposition stages should be included in the training data. Therefore, we randomly selected a small sample of subjects (or donors) and then a subset of images, of different body parts, taken from them over time (i.e., from when they first started to decay until fully decomposed) instead of randomly selecting images from the entire human decomposition dataset. This resulted in a subset of 4731 images of multiple body parts across different stages of decay, which were labeled by a forensic expert with a body part label (including head, torso, arm, hand, leg, or foot) and a decay score (DS) label, ranging from 1 (fresh) to 13 (fully skeletonized), using a modification of Galloway’s [16] human decomposition scoring method to indicate the amount of decay present. The labeled dataset was named SHD: Stages of Human Decomposition. Figures 3.2 and 3.3 show sample images of SHD with image resolutions varying from 2400×1600 up to 4900×3200 . Since the aim is to predict the stage or level of decay rather than the DS itself, each image’s DS label was mapped into DS intervals, carefully defined by the forensic experts, representing a certain SOD. As advised by the forensic experts, 3-4 stages are commonly used to describe the human decay process. Therefore, two versions of SHD were created by mapping each image’s DS label into each of the following two sets of DS intervals: (1) 1-4 (fresh), 5-9 (active decay), and 10-13 (skeletonization), and (2) 1-4 (fresh), 5-7 (early decay), 8-9 (advanced decay), and 10-13 (skeletonization). The mapped versions of SHD, named SHD-3 (for 3 SOD classes) and SHD-4 (for 4 SOD classes), along with their DS groups (or intervals), class names, and frequencies (N) are shown in Table 3.1.



Figure 3.2: No decay. The image has been blurred to make them less disturbing to sensitive viewers.



Figure 3.3: Fully decayed. The image has been blurred to make them less disturbing to sensitive viewers.

Table 3.1: The SHD-3 and SHD-4 datasets along with their DS groups, class names, and class frequencies (N). Note, the class active decay = early + advanced decay.

Dataset	DS group	Class name	N
SHD-3	1-4	fresh	1479
	5-9	active decay	2053
	10-13	skeletonization	1199
SHD-4	1-4	fresh	1479
	5-7	early decay	1095
	8-9	advanced decay	958
	10-13	skeletonization	1199

3.3.3 Proposed Label Propagation Method

The developed label propagation method leverages exogenous attributes, including the donor ID and timestamp (i.e., the date the image was taken), and endogenous attributes, including the body part shown in the image, to propagate the SOD labels of the labeled samples to the unlabeled samples. The image donor ID and timestamp are known attributes contained in the image filename, but the body part is unknown and is derived prior to label propagation using an existing body part classifier. This algorithm was named ExoEndo-LP for Exogenous-Endogenous-Based Label Propagation. ExoEndo-LP accepts the parameters, (1) *id* (image donor ID), (2) *time* (image timestamp) and (3) *bp* (image body part), indicating the exogenous and endogenous attributes the algorithm considers during label propagation, which align with the attributes considered by the forensic experts during manual labeling. Note, the *id* parameter is a required parameter, while the *time* and *bp* are optional parameters. Specifically, ExoEndo-LP works as follows: let $I = \{(x_i, (b_i, s_i), d_i, t_i)\}_{i=1}^n$ be a manually labeled dataset with n images, where (b_i, s_i) are the body part and SOD labels, respectively, d_i the donor ID, and t_i the image timestamp for the i^{th} labeled image, x_i . If $\text{ExoEndo-LP}(id, time)$, meaning label propagation is subject and time-specific, then given a labeled image, $(x_i, (b_i, s_i), d_i, t_i) \in I$, all images $m \notin I$ of donor, d_i , with the timestamp, t_i , are found and denoted as the set

$S_i = \{(x_j, d_i, t_i)\}_{j=1}^m$. Then, the SOD label, s_i , of the labeled image, x_i , is propagated throughout S_i resulting in the labeled set $L_i = \{(x_j, s_i, d_i, t_i)\}_{j=1}^m$. Else if $\text{ExoEndo-LP}(id, time, bp)$, meaning label propagation is subject, time, and body part-specific, then given a labeled image, $(x_i, (b_i, s_i), d_i, t_i) \in I$, all images $m \notin I$ of donor, d_i , with the timestamp, t_i , are found and denoted as the set $S_i = \{(x_j, d_i, t_i)\}_{j=1}^m$. Since the bp parameter was passed in addition to the id and $time$ parameters, only images in S_i of the same body part as the labeled image, x_i , will be labeled with s_i . Therefore, a previously developed body part classifier (with a test macro-averaged F1 score of 0.93 and a test accuracy of 0.94) was used to automate this task by classifying each image in S_i into a body part (see Section 3.3.2 for body part labels), resulting in the set of l images denoted as $S_{i+} = \{(x_k, \hat{b}_k, d_i, t_i)\}_{k=1}^l$ where $\hat{b}_k$ is the predicted body part such that $\hat{b}_k = b_i$ for the k^{th} image, x_k . Then, the SOD label, s_i , of the labeled image, x_i , is propagated throughout S_{i+} , resulting in the labeled set $L_i = \{(x_k, (b_i, s_i), d_i, t_i)\}_{k=1}^l$. In either case, the labeled set L_i gets added to a final set of labeled images, $I_+ = I \cup \{L_i\}_{i=1}^n$.

To evaluate the performance of the proposed ExoEndo-LP, the standard graph-based label propagation algorithm (LPA) [48], was used as a baseline method. Unlike ExoEndo-LP, LPA [48] does not consider any image attributes, but is based on the assumption that closer data points tend to have similar class labels. Specifically, at every iteration of propagation, each unlabeled sample is labeled with the label that majority of its neighbors belong to or to a random selection when there is no single maximum. LPA [48] reaches convergence when each unlabeled sample has the majority label of its neighbors, or the predefined maximum number of iterations is achieved. To reduce the size of each labeled and unlabeled image, a previously trained classifier, excluding the last fully-connected and soft max layers, was used. Specifically, the output vectors of length 256 were used as the image vector representations to which LPA [48] was then applied.

3.3.4 SOD Classification

To build the SOD classifiers, transfer learning was applied, which aims to produce effective models by leveraging and exploiting previously acquired knowledge [45]. In particular, three CNN architectures, including ResNet50 [18], Inception V3 [55], and Xception [6], pre-trained on the ImageNet [11] dataset, were trained using the following two-step transfer learning process: (1) freeze all pretrained convolutional layers of the base model, and train newly added classifier layers and (2) unfreeze all layers, and fine-tune model end-to-end with a low learning rate. The newly added classifier layers consisted of five layers, including one global average pooling layer, one drop-out of 0.3 layer, two fully-connected layers with 128 and 64 nodes, respectively, and one final softmax layer with the number of nodes equal to the number of classes to perform multi-class classification. In addition, to increase diversity during training, a data augmentation layer, which performed image flipping (horizontal and vertical) and rotating, was added after the input layer. Adaptive Momentum Estimation (Adam) was used as the optimizer with a learning rate of 0.001 and 0.0001 for the first and second step in the two-step transfer learning process, respectively. Note, training a model from scratch and transfer learning without freezing the base model, and freezing only a certain number of base model layers followed by fine-tuning was also tested, however, model performances were drastically improved using the proposed two-step transfer learning process.

3.3.5 Implementation and Experimental Settings

The developed SOD-VS was implemented using Keras, TensorFlow, and Python, and MongoDB as the database. The SOD-VS was built and evaluated on the manually labeled datasets, SHD-3 and SHD-4, using the proposed ExoEndo-LP and baseline, LPA [48], methods to perform label propagation. Note, label propagation was only performed on the train and validation data, but not on the test data to ensure equal classification evaluation. To evaluate the proposed ExoEndo-LP, ExoEndo-LP(id,

time) and ExoEndo-LP(id, time, bp) were tested. The baseline method, LPA [48], was implemented using 1000 maximum iterations and a varying number of neighbors (i.e., 5, 10, and 20). Therefore, a total of 30 experiments (2 datasets $\times$ (3 LPA [48] + 2 ExoEndo-LP) $\times$ 3 CNN architectures) were conducted. To train and evaluate the SOD classifiers, each labeled dataset (SHD-3 and SHD-4) was split into a train, validation, and test set using a ratio of 70:10:20, resulting in 3312 train samples, 473 validation samples, and 946 test samples. All images were resized to 224x224 and 299x299 depending on the model architecture. All models were trained on a single TeslaV100-SXM2 GPU with 32GB memory. The batch size was set to 32 and the number of epochs to 200 with early stopping set to 20 epochs (monitoring the validation loss) to avoid over-fitting on the training set. To evaluate the SOD classifiers, the per-class precision and recall, and macro-averaged F1 score (mF1) on the test data were calculated.

3.4 Experimental Results and Discussion

The average runtimes and number of additionally labeled samples after the baseline method, LPA [48], and the proposed ExoEndo-LP were performed on the train data are reported in Table 3.2, which shows that the proposed ExoEndo-LP was up to twice as fast as the baseline method, LPA [48]. The baseline method, LPA [48], labeled more samples, which aligns with the fact that since ExoEndo-LP considers image attributes, it is more restrictive, meaning less candidates are considered during labeling, resulting in less labeled images than with LPA [48] that labels each unlabeled image based on the majority class of its neighbors.

Although, ExoEndo-LP produced a smaller labeled train set, higher SOD classification performances were achieved as shown by Table 3.3. In fact, SOD-VS+ExoEndo-LP(id, time, bp) using Inception-V3 trained on SHD-3 improved the mF1 score from 0.79 (achieved by SOD-VS+LPA [48] using Inception-V3) to 0.863 (9.24%), and SOD-VS+ExoEndo-LP(id,time,bp) using Xception trained on SHD-4

Table 3.2: Performance comparison of the baseline method, LPA [48], and the proposed method, ExoEndo-LP, when applied to the train data (n=3312). Reported are average runtimes (in seconds, s) and the number of additionally labeled (NoL) train samples using label propagation.

Method	Runtime (s)	NoL
LPA [48]	32	14067
ExoEndo-LP(<i>id, time</i>)	16	6034
ExoEndo-LP(<i>id, time, bp</i>)	20	2250

improved the mF1 score from 0.695 (achieved by SOD-VS+LPA [48] using Xception) to 0.803 (15.54%). This suggests that our method produced synthetic labels of quality more similar to that of the manual labels produced by the forensic experts, instead of more numerous but, presumably, less accurate synthetic labels produced by the baseline label propagation method. Note, SOD-VS+LPA [48] performed best using 5 neighbors so only these results are reported in Table 3.3.

For the three and four class case, classification performance was improved when the body part in addition to donor ID and timestamp was considered during label propagation (i.e., ExoEndo-LP(*id, time, bp*)). Since the value of improvement for mF1 from 0.83 to 0.863 in the three-class case and from 0.78 to 0.803 in the four-class case was relatively small we tested the statistical significance of this improvement via a McNemar’s test [31] and found both to be statistically significant at 0.005 level.

Overall, classification results were higher when three SOD classes were used, meaning when early and advanced decay were combined which according to the forensic experts are the most difficult to differentiate between. Also, ResNet-50 was outperformed in all experiments, which could be due to its deep architecture requiring longer training time than more shallow architectures, such as Inception-V3 and Xception.

Table 3.3: SOD-VS with the baseline method, LPA [48] (neighbors=5), and the proposed ExoEndo-LP evaluated on the SHD-3 test data (top three) and SHD-4 test data (bottom three).

Method	Model	Per-Class Precision (P) & Recall (R)								mF1
		fresh		early/active		advanced		skeletonization		
		P	R	P	R	P	R	P	R	
SOD-VS+LPA [48]	ResNet-50	0.0	0.0	0.603	0.8	–	–	0.703	0.629	0.45
	Inception-V3	0.713	0.682	0.851	0.804	–	–	0.827	0.853	0.79
	Xception	0.7	0.554	0.819	0.844	–	–	0.834	0.844	0.763
SOD-VS+ExoEndo-LP(<i>id, time</i>)	ResNet-50	0.77	0.555	0.396	0.22	–	–	0.551	0.91	0.54
	Inception-V3	0.869	0.844	0.723	0.863	–	–	0.824	0.861	0.83
	Xception	0.814	0.85	0.718	0.72	–	–	0.812	0.778	0.78
SOD-VS+ExoEndo-LP(<i>id, time, bp</i>)	ResNet-50	0.797	0.441	0.522	0.911	–	–	0.771	0.157	0.497
	Inception-V3	0.927	0.844	0.813	0.906	–	–	0.892	0.809	0.863
	Xception	0.86	0.944	0.869	0.81	–	–	0.807	0.817	0.85
SOD-VS+LPA [48]	ResNet-50	0.5	0.068	0.67	0.324	0.0	0.0	0.568	0.944	0.318
	Inception-V3	0.757	0.818	0.701	0.64	0.627	0.576	0.709	0.629	0.683
	Xception	0.767	0.659	0.81	0.83	0.567	0.222	0.96	0.902	0.695
SOD-VS+ExoEndo-LP(<i>id, time</i>)	ResNet-50	0.707	0.691	0.294	0.12	0.0	0.0	0.54	0.9	0.388
	Inception-V3	0.821	0.922	0.64	0.732	0.705	0.505	0.876	0.779	0.74
	Xception	0.91	0.954	0.696	0.771	0.71	0.468	0.911	0.884	0.78
SOD-VS+ExoEndo-LP(<i>id, time, bp</i>)	ResNet-50	0.567	0.801	0.349	0.494	0.288	0.108	0.631	0.277	0.403
	Inception-V3	0.844	0.916	0.706	0.705	0.642	0.675	0.923	0.766	0.773
	Xception	0.897	0.908	0.734	0.8	0.75	0.619	0.851	0.872	0.803

3.5 Conclusion

This paper presents a novel vision system that utilizes a large human decomposition image dataset to perform SOD classification. The system uses a label propagation method that leverages the exogenous and endogenous image attributes with the constraints based on domain knowledge to improve classification performance (and reduce the manual labeling efforts). The approach is general in the sense that image collections in other domains tend to have exogenous and endogenous attributes that could be used in a manner similar to the one employed in this work.

The improvements from the basic label propagation make the classification sufficiently accurate for a number of downstream applications, such as post-mortem interval estimation. Future work will focus on improving the classification performances of the early and advanced decay classes since all CNN architectures struggled to distinguish the two to improve the accuracy of the four-SOD-class case.

We hope this work can help progress the research of forensic anthropology in conjunction with computer vision and be used in other domains where, similar to human decay, a large degree of variation exists and human labeling is limited from both a time and resource-perspective.

Chapter 4

Towards Automation of Human Stage of Decay Identification: An Artificial Intelligence Approach

Disclosure Statement

This chapter has been submitted to a journal for review and as a preprint [40].

My contributions: I was responsible for the investigation, methodology, development, and writing. Specifically, I managed data extraction and processing for model development and the interrater study, as well as model training, testing, and results analysis. This also included efforts related to reporting and publication.

4.1 Abstract

Determining the stage of decomposition (SOD) is crucial for estimating the post-mortem interval and identifying human remains. Currently, labor-intensive manual scoring methods are used for this purpose, but they are subjective and do not scale for the emerging large-scale archival collections of human decomposition photos. This study explores the feasibility of automating two common human decomposition scoring methods proposed by Megyesi and Gelderman using artificial intelligence (AI). We evaluated two popular deep learning models, Inception V3 and Xception, by training them on a large dataset of human decomposition images to classify the SOD for different anatomical regions, including the head, torso, and limbs. Additionally, an interrater study was conducted to assess the reliability of the AI models compared to human forensic examiners for SOD identification. The Xception model achieved the best classification performance, with macro-averaged F1 scores of .878, .881, and .702 for the head, torso, and limbs when predicting Megyesi’s SODs, and .872, .875, and .76 for the head, torso, and limbs when predicting Gelderman’s SODs. The interrater study results supported AI’s ability to determine the SOD at a reliability level comparable to a human expert. This work demonstrates the potential of AI models trained on a large dataset of human decomposition images to automate SOD identification.

4.2 Introduction

Determining the stage of decay (SOD) is an important and common task in human remains cases. Knowing the degree of decomposition is vital for estimating the post-mortem interval (PMI) and identifying human remains [33, 17, 57, 16]. Presently, establishing the SOD of a decedent is primarily conducted manually, via visual assessment by trained experts using non-metric scoring or staging methods, such as those proposed by Megyesi et al. [33] and Gelderman et al. [17]. Such non-metric methods, which rely on subjective interpretation made by humans, possess a higher susceptibility to human bias and error, consequently affecting the accuracy of downstream tasks, such as estimating the PMI [38, 10, 25, 51, 13]. Furthermore, the PMI estimation formulas derived in existing studies, such as those by Megyesi et al. [33] and Gelderman et al. [17], were developed using a very small number of samples. Evaluating or improving upon these formulas with a much larger sample size, such as over one million photos, would require manual SOD scoring, which is not feasible for such a large sample. Therefore, this work aims to utilize emerging artificial intelligence (AI) methods to evaluate the feasibility of automating the SOD identification task.

AI is the ability of machines to perform tasks that would typically require human intelligence and it has provided innovative approaches to assist in human decision-making [12, 21]. AI assesses information based on the entirety of acquired facts or data using advanced algorithms, thereby mitigating vulnerability to the subjectivity and biases that trouble humans and affect their decision-making abilities [47]. Additionally, AI algorithms can handle large volumes of data, uncovering intricate patterns that might elude human perception [12, 49, 24]. This ability to process, analyze, and interpret large amounts of data quickly and precisely makes AI a valuable tool in many industries, including forensic practice and research.

In summary, the objective of this study is to evaluate the possibility of automating two established human decomposition scoring methods, namely Megyesi et al. [33]

and Gelderman et al. [17], using vision-based AI models, known as convolutional neural networks (CNNs). Specifically, various CNN classification models will be trained and evaluated on a large human decomposition image dataset to perform SOD prediction. In addition, an interrater test is conducted to assess and compare the reliability of the models and the human forensic examiners for SOD identification. We hypothesize that similar interrater reliability among human raters and an AI rater suggests the feasibility of using AI for SOD classification and, perhaps, other downstream tasks. The significance of such a finding lies in the potential to develop more accurate SOD and PMI estimation methods that are less effort-intensive and subjective. The primary purpose of this study is to provide a proof-of-concept for the future advancement and integration of AI-assisted analysis in forensic practice and research.

4.3 Materials and Methods

4.3.1 Stage of Decay Scoring Methods

The two human decomposition scoring methods this study attempts to automate using AI are: (1) Megyesi et al. [33] (see Appendices B.1 to B.3) and (2) Gelderman et al. [17] (Appendices B.4 to B.6). To account for the differential decomposition that occurs in different body segments (e.g., limbs do not bloat or purge fluid), these two scoring methods independently assess the human body in three anatomical regions: (1) the head (including the neck), (2) the torso, and (3) the limbs (including the hands and feet). Based on the morphological features present, Megyesi et al. [33] categorizes human decomposition into four high-level linear stages: fresh, early decomposition, advanced decomposition, and skeletonization. Gelderman et al. [17], building upon the work of Megyesi et al. [33], categorizes each anatomical region into six stages, with the lowest indicating no visible changes and the highest indicating complete skeletonization.

4.3.2 The Human Decomposition Dataset

The human decomposition dataset, a large-scale image collection used to train the models, includes images of decomposing donors donated to the Forensic Anthropology Center at The University of Tennessee, Knoxville. The center houses the Anthropology Research Facility (ARF), an outdoor decomposition laboratory. Forensic experts from the ARF captured these images at non-uniform intervals, with one or more days between each capture. The images, taken from various angles, depict different anatomical areas to illustrate the various stages and regions of human decomposition. The image resolutions vary from 2400×1600 pixels up to 4900×3200 pixels. The dataset covers the period from 2011 to 2023 and comprises over 1.5 million images contributed by more than 800 donors. To train different CNN classifiers on this large human decomposition image dataset to predict the SOD for various anatomical regions, the following challenges needed to be addressed:

- How to best sample from the entire human decomposition dataset such that the resulting set of images reflects the dataset’s temporal characteristics. In other words, the data used to train the models should consist of images covering the entire human decay process, that is, from death to skeletonization.
- Once a set of images has been sampled, the challenge is to automate the efficient extraction of specific anatomical regions (i.e., head, torso, and limbs). From a time- and cost-effective perspective, it is not feasible to manually perform this body part filtering of over one million images.

The following section (Section [4.3.3](#)) details how these analysis challenges were addressed during data preparation.

4.3.3 Data Processing and Labeling

The human decomposition dataset was processed according to the data pipeline shown in Figure 4.1. The remainder of this section further discusses the individual steps of this pipeline.

The quality and size of the data used to train the models highly affect a model’s performance and generalizability. In other words, the more representative and diverse the training data is, the more likely it is that the model will be able to generalize. In the case of a temporal dataset, such as images documenting human decomposition, where the subjects’ appearance changes over time, it is important to sample the training data in a way that reflects the dataset’s characteristics. Therefore, images depicting all possible decomposition stages should be included in the training data. As a result, a small sample of donors, and all images of those donors, was selected over time (i.e., from when they first started to decay until fully decomposed), instead of randomly selecting images from the entire human decomposition dataset. Since this subset of images included all different kinds of anatomical areas, a previously developed body part classification model was used to automatically detect the head, torso, and limbs images to align with Megyesi et al.’s [33] and Gelderman et al.’s [17] scoring methods. The next step was to manually label this subset of images with the SOD labels. Specifically, the two scoring methods, described in Section 4.3.1, were used by a forensic expert to perform manual data labeling of the head, torso, and limbs images. Note that any body part misclassifications and/or poor quality images were either corrected or removed at this point to ensure that the final labeled datasets only included high quality images of the head, torso, and limbs. Table 4.1 shows the SOD terms used in the original literature (i.e., Megyesi et al. [33] and Gelderman

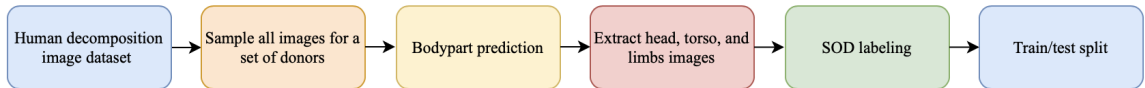


Figure 4.1: The data processing pipeline.

Table 4.1: The original SOD terms mapped to the new SOD class labels used throughout this study for the different scoring methods.

Method	Original SOD Term	New SOD Class Label
Megyesi et al. [33]	fresh	M-SOD1
	early decomposition	M-SOD2
	advanced decomposition	M-SOD3
	skeletonization	M-SOD4
Gelderman et al. [17]	1	G-SOD1
	2	G-SOD2
	3	G-SOD3
	4	G-SOD4
	5	G-SOD5
	6	G-SOD6

et al. [17]) and the corresponding new class labels used throughout this study. For instance, if the forensic expert, applying Megyesi et al.’s [33] method, determined the SOD of a head image to be *fresh*, then it was assigned the *M-SOD1* class label.

The image labeling was conducted using an in-house developed data visualization and annotation software called ICPUTRD (Image Cloud Platform for Use in Tagging and Research on Decomposition) [42]. The resulting labeled datasets are shown in Table 4.2, each of which will be used to train a SOD classification model. For example, *M-head-data* will be used to train a model to predict the SOD (i.e., *M-SOD1*, *M-SOD2*, *M-SOD3*, or *M-SOD4*) of head images. Finally, each labeled dataset was split into a train and test set using a ratio of 80:20. The train set was used to train the model and the test set was used to evaluate the model once trained.

4.3.4 AI Model Development

To build the SOD classifiers, transfer learning was applied, which aims to produce effective models by leveraging and exploiting previously acquired knowledge [45]. In particular, two CNN architectures, including Inception V3 [55] and Xception [6], pre-trained on the ImageNet dataset [11], were trained using the following two-step

Table 4.2: The labeled datasets used to train the SOD classifier models. Method indicates the SOD scoring method used to label the different anatomical regions. NoL gives the number of labeled images for the particular dataset.

Method	Anatomical Region	Dataset	NoL
Megyesi et al. [33]	head	M-head-data	2110
	torso	M-torso-data	1979
	limbs	M-limbs-data	2152
Gelderman et al. [17]	head	G-head-data	2041
	torso	G-torso-data	1982
	limbs	G-limbs-data	2032

transfer learning process: (1) freeze all pre-trained convolutional layers of the base model and train newly added classifier layers and (2) unfreeze all layers and fine-tune the model end-to-end with a low learning rate. The newly added classifier layers consisted of five layers, including one global average pooling layer and one drop-out layer (rate = 0.3) to alleviate the over-fitting problem motivated by Lin et al. [27], followed by two fully-connected layers with 128 and 64 nodes performing down-sampling, and one final softmax layer with the number of nodes equal to the number of classes for multi-class classification. More precisely, the softmax layer transforms the output of the previous layer into a probability distribution over all the classes as shown by the equation (4.1), with the class having the highest probability being the final SOD prediction. In equation (4.1), z_i is the i^{th} element of the input vector to the softmax function and K the total number of classes. To increase the size and diversity of the data, a data augmentation layer was added after the input layer, performing random image flipping (horizontal and vertical) and rotation during model training. Figure 4.2 gives an overview of the developed SOD classification framework.

$$\text{softmax}(z_i) = \frac{e^{z_i}}{\sum_{j=1}^K e^{z_j}} \quad \text{for } i = 1, 2, \dots, K \quad (4.1)$$

The loss function used was Cross-Entropy loss, which takes the predicted probability distribution of the softmax layer and measures how well this distribution

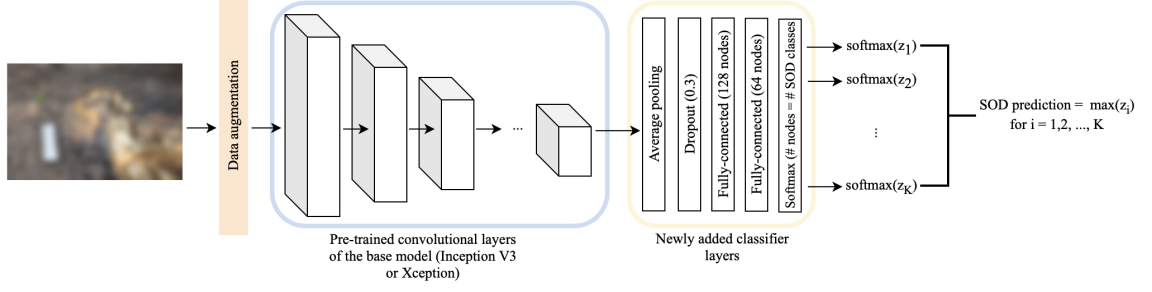


Figure 4.2: The SOD classification framework was trained using a two-step transfer learning approach. In step 1, the pre-trained convolutional layers of the base model (blue rectangle) were frozen, and only the newly added classifier layers (yellow rectangle) were trained. In step 2, the previously frozen layers from step 1 (blue rectangle) were unfrozen, and the entire model was trained end-to-end.

matches the true distribution. To minimize the Cross-Entropy loss function during training, the Adaptive Momentum Estimation (Adam) optimizer was employed, with a learning rate of 0.001 for the first step and 0.0001 for the second step in the two-step transfer learning process. It is worth noting that training the Inception V3 and Xception architecture from scratch (i.e., without transfer learning), transfer learning without freezing the base model (i.e., one-step transfer learning), and freezing only a certain number of base model layers instead of all followed by fine-tuning was also tested. However, the proposed two-step transfer learning process significantly improved model performances.

To evaluate the performance of the trained SOD classifier models, the confusion matrix on the test data was calculated per class, which summarizes a model's performance by comparing its predicted labels to its true labels. Specifically, the confusion matrix shows the number of correct predictions, such as the true positives (TP) and the true negatives (TN), as well as the number of incorrect predictions, such as the false positives (FP) and the false negatives (FN). The following two performance metrics were then calculated from the per-class confusion matrix: precision (4.2) and recall (4.3). Precision measures the accuracy of the positive predictions, while recall measures the completeness of the positive predictions.

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (4.2)$$

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (4.3)$$

To combine the per-class precision and recall metrics into a single model evaluation metric, the macro-averaged F1 score (mF1) (4.4) was calculated and reported, which is the unweighted mean of the per-class F1 scores (4.5).

$$\text{mF1} = \frac{\sum_{i=1}^n \text{F1}_i}{n} \quad \text{for } i = 1, 2, \dots, n, \text{ where } n \text{ is the number of classes} \quad (4.4)$$

$$\text{F1} = \frac{2 * (\text{Precision} * \text{Recall})}{\text{Precision} + \text{Recall}} \quad (4.5)$$

4.3.5 Interrater Test

To ensure that manual rating techniques are reliable, interrater reliability tests are often used to assess how similar the ratings are between two or more raters on the same set of data samples. In this study, it is also used to assess the reliability of the developed models. The interrater test involved multiple raters (both human and AI) labeling a set of images depicting the same anatomical region using Megyesi et al.’s [33] and Gelderman et al.’s [17] scoring methods. Specifically, 300 torso images that were not used during model development were selected. Due to limited resources, the interrater test focused on the torso only, which provides a good amount of variability. The raters included the developed torso SOD model or *Model* and three forensic experts well-versed in the considered scoring methods, including *Human 1* (the same human who labeled the data the models were trained with), *Human 2*, and *Human 3*. The task for each rater was to label the 300 images using once Megyesi et al.’s [33] method and once Gelderman et al.’s [17] method. The *Model* rater performed labeling

by predicting the SOD of the 300 images, while the human raters were instructed to independently label the 300 images on ICPUTRD [42], following a similar set-up used for data labeling in Section 4.3.3. To ensure randomization across methods, human labeling was conducted in batches of images instead of labeling all 300 images with one method and then the other. Specifically, the human raters were presented with batches of 50 images at a time, which they were asked to label using one scoring method (i.e., Megyesi et al. [33] or Gelderman et al. [17]). This process was repeated until all 300 images were labeled with both scoring methods, resulting in 12 iterations for each human rater $((300 \text{ images} / 50 \text{ images}) \times 2 \text{ methods})$. Note, the method used to label a batch of images was alternated between the two scoring methods.

After each rater completed labeling, two types of agreements were assessed: (1) human-human agreement (i.e., the agreement among all three human raters: *Human 1*, *Human 2*, and *Human 3*) and (2) AI-human agreement (i.e., the agreement among the *Model* replacing *Human 1*, and the other two human raters, *Human 2*, and *Human 3*). The reason for the *Model* replacing *Human 1* was to see how the agreement changes when the human is replaced by the model trained on the data they labeled. To measure the different agreements, the Fleiss’ Kappa statistic was used, which measures the degree of agreement among raters over what would be expected by random chance, with values ranging from -1 (no agreement) to 1 (perfect agreement). The Fleiss’ Kappa values will be interpreted based on Landis and Koch’s [26] interpretation criteria shown in Table 4.3.

Table 4.3: Landis and Koch’s [26] Fleiss’ Kappa interpretation criteria.

Fleiss’ Kappa	Level of Agreement
≥ 0.8	almost perfect
≥ 0.6	substantial
≥ 0.4	moderate
≥ 0.2	fair
≥ 0	slight
< 0	no agreement

4.4 Results

4.4.1 SOD Classification

The SOD classifier models were implemented using Keras and TensorFlow, two open-source machine learning modules written in Python. In particular, the models were built and evaluated on the six manually labeled datasets shown in Table 4.2. All images were resized to 299×299 pixels, as that is the required input image size for both the Inception V3 and Xception architectures. All models were trained on a single Tesla V100-SXM2 GPU with 32GB of memory. The batch size was set to 32, and the number of epochs was set to 200, with early stopping set to 20 epochs to avoid over-fitting on the training set.

Table 4.4 presents the Megyesi et al. [33] SOD classification results on the test data. For each anatomical region, the best SOD classification performance, as indicated by the mF1 scores, was achieved with the Xception architecture. The head SOD classifier achieved an mF1 score of .878, and the torso SOD classifier achieved an mF1 score of .881, performing comparably. However, the limbs SOD classifier performed notably less well, with an mF1 score of .702.

Table 4.5 presents the Gelderman et al. [17] SOD classification results on the test data. Similar to the Megyesi et al. [33] results, the best classification performance,

Table 4.4: Megyesi et al.’s [33] SOD classification results.

Archt.	Region	Per-Class Precision (P) & Recall (R)								mF1
		M-SOD1		M-SOD2		M-SOD3		M-SOD4		
		P	R	P	R	P	R	P	R	
InceptionV3	head	.846	.805	.835	.91	.698	.677	.882	.778	.806
	torso	1.0	1.0	.936	.83	.683	.707	.733	.892	.845
	limbs	.667	.333	.94	.94	.467	.5	.897	.929	.695
Xception	head	.932	.842	.878	.966	.887	.723	.893	.926	.878
	torso	1.0	1.0	.926	.898	.754	.767	.829	.872	.881
	limbs	.75	.5	.924	.978	.6	.214	.92	.967	.702

Table 4.5: Gelderman et al.’s [17] SOD classification results.

Archt.	Region	Per-Class Precision (P) & Recall (R)												mF1
		G-SOD1		G-SOD2		G-SOD3		G-SOD4		G-SOD5		G-SOD6		
		P	R	P	R	P	R	P	R	P	R	P	R	
InceptionV3	head	1.0	.8	.903	.933	.946	.907	.8	.8	.696	.8	.895	.944	.866
	torso	.2	.5	.862	.862	.954	.954	.897	.833	.7	.824	.833	.714	.749
	limbs	.75	.75	.704	1.0	.907	.739	.696	.765	.512	1.0	1.0	0.059	.651
Xception	head	1.0	.6	.882	1.0	.947	.918	.806	.829	.818	.9	1.0	.889	.872
	torso	.667	1.0	.964	.931	.974	.862	.792	1.0	.789	.81	.944	.81	.875
	limbs	1.0	1.0	.875	.947	.909	.87	.75	.824	.826	.905	1.0	0.071	.76

as indicated by the mF1 scores, was achieved by the Xception architecture across all anatomical regions. The head SOD classifier achieved an mF1 score of .872, and the torso SOD classifier achieved an mF1 score of .875, performing comparably. However, the limbs SOD classifier performed less well, with an mF1 score of .76.

4.4.2 Interrater Test

The interrater test results are shown in Table 4.6. Specifically, a Fleiss’ Kappa analysis using SPSS Statistics was conducted for both the human-human agreement and AI-human agreement across both scoring methods. The reported agreement levels were determined using Landis and Koch’s [26] interpretation criteria, as shown in Table 4.3. According to the Megyesi et al. [33] results, the Fleiss’ Kappa coefficient of the human-human agreement was .67 with a p-value < .001 and a 95% confidence interval (CI) of .628 to .713, indicating *substantial* agreement. Similarly, the Fleiss’ Kappa coefficient of the AI-human agreement was .637 with a p-value < .001 and a 95% CI of .594 to .68, suggesting *substantial* agreement. Additionally, according to the Gelderman et al. [17] results, the Fleiss’ Kappa coefficient of the human-human agreement was .593 with a p-value < .001 and a 95% CI of .558 to .628, indicating *moderate* agreement. Similarly, the Fleiss’ Kappa coefficient of the AI-human agreement was .558 with a p-value < .001 and a 95% CI of .524 to .592, suggesting *moderate* agreement. In all

Table 4.6: The interrater study results. Reported are the Fleiss’ kappa value, the p-value, the upper bound (UB) and lower bound (LB) of the 95% confidence interval (CI), and the agreement level for both the human-human and AI-human agreement across both scoring methods. The agreement levels are determined using Landis and Koch’s [26] interpretation criteria, as shown in Table 4.3.

Method	Agreement	Fleiss’ Kappa	P-value	95% CI		Agreement Level
				UB	LB	
Megyesi et al. [33]	human-human	.67	<.001	.628	.713	substantial
	AI-human	.637	<.001	.594	.68	substantial
Gelderman et al. [17]	human-human	.593	<.001	.558	.628	moderate
	AI-human	.558	<.001	.524	.592	moderate

cases, the Fleiss’ Kappa coefficient was statistically significant (i.e., the p-value < 0.05).

4.4.3 Discussion

Overall, the SOD classification results are promising. The Xception architecture performed the best across both scoring methods. The head and torso SOD models performed comparably well; however, the limbs SOD models’ performances were not as strong. Further analysis of the limb data revealed that some images included hands and/or feet covered by a net (to prevent them from being scattered/disarticulated by animal scavengers), which could confuse and distract the model, leading to incorrect predictions. Additionally, the examination of the limbs data indicated that some images included other parts of the body, specifically the torso. Since the torso decays differently than the limbs, this inclusion could again confuse the model and, consequently, affect its predictions. Future work will focus on addressing these data quality challenges to ensure the development of a more reliable limbs dataset and hence improved prediction performance. Another important finding to mention is that although the Gelderman et al. [17] datasets contained more SOD classes (six classes) than the Megyesi et al. [33] datasets (four classes), the classification performances

were comparable. This indicates that these AI models are able to learn a higher number of decay stages without decreasing prediction performance.

While the SOD prediction performances were promising overall, there is room for improvement, as indicated by the per-class precision and recall values. The sizes of the labeled datasets used in this study are considered rather small for training deep learning architectures, such as Inception V3 and Xception. Additionally, a deeper analysis of the labeled datasets indicated class imbalance (i.e., a disproportionate number of instances of one class compared to another). Having a larger, more diverse, and evenly balanced dataset will make the models more robust and improve their generalization capabilities, which refers to how well a model can react to new and unseen data. However, creating more labeled data where domain expertise is required may be limited by both resource and time constraints. A recent study by Nau et al. [41] addressed such challenges by developing a domain-aware label propagation algorithm that leverages different image attributes to automatically perform data labeling, thereby reducing manual labeling efforts and costs. Future work will explore integrating such label propagation methods to obtain larger and more diverse datasets, aiming to create more robust and accurate SOD classification models.

In the interrater test, both the human-human and AI-human agreements showed *substantial* agreement when applying the Megyesi et al. [33] scoring method. Using the Gelderman et al. [17] scoring method, *moderate* agreement was observed for both the human-human and AI-human agreements. Across both scoring methods, the level of agreement for both the human-human and AI-human agreements was the same. This means that when the human rater was replaced with the AI model rater, the level of reliability stayed the same, supporting AI’s ability to perform SOD identification with a reliability level comparable to that of an experienced human forensic examiner. Notably, the lower agreement level for the Gelderman et al. [17] method could be attributed to (1) its novelty in the field and/or (2) its complexity,

being more complicated to apply, as it involves six decay stages compared to the four stages in the Megyesi et al. [33] method.

While the results demonstrate that human decomposition scoring methods have the potential to be automated using AI techniques, there are some important limitations of this work. For one, the data labeling was conducted by a single forensic expert. This approach may introduce labeling bias, which can lead to inherently biased training datasets. Models trained on such datasets can inherit these biases, resulting in biased models [22]. Therefore, future work will focus on creating a so-called “gold standard” dataset—a labeled dataset meticulously crafted and evaluated by multiple forensic experts. Such a dataset would be accepted as the most accurate and reliable of its kind. This step will ensure that the models are trained with accurate and unbiased data, which is vital for developing high-quality models.

An additional limitation is that this study is environmental- and climate-specific. The images used to train the models were all taken of donors decaying outdoors in an open-wooded area with the ground consisting of soil, gravel, and dead/decaying plant matter (e.g., rotting wood and shedding leaves). Additionally, the climate of this area is humid subtropical, characterized by high summer and moderate winter temperatures. Therefore, the models may not perform as well on images taken in different climate conditions or environments, necessitating additional training or re-training with images specific to those conditions.

4.5 Conclusion

This study explored the possibility of automating two common human decomposition scoring methods, namely Megyesi et al. [33] and Gelderman et al. [17]. Specifically, different CNN models, including Inception V3 and Xception, were trained on a large human decomposition image dataset to classify the SOD for different anatomical regions. Across both scoring methods, the Xception model achieved the highest classification results, performing comparably well for the head and torso, and slightly

lower for the limbs. The interrater reliability study results provided support for AI's ability to automate the SOD identification task at a reliability level comparable to a human expert. Overall, the study results are promising and provide a proof-of-concept for automating human decomposition scoring methods using AI.

Chapter 5

Identifying Factors to Help Improve Existing Decomposition-Based PMI Estimation Methods

Disclosure Statement

This chapter has been submitted to a journal for review and as a preprint [39].

My contributions: Investigation, methodology, development, and writing. Specifically, I was in charge of data extraction and processing, statistical modeling, and results analysis and reporting, including efforts related to publication.

5.1 Abstract

Accurately assessing the post-mortem interval (PMI), or the time since death is an important but challenging task in forensic science. Existing regression model formulas use the decomposition score to predict the PMI or accumulated degree days (ADD), but the provided formulas are based on small samples and lack precision. We aim to explore if (a) with much larger samples, (b) more sophisticated statistical models, and (c) a richer set of predictors derived from demographic and environmental factors known to affect the human decay process, we can construct substantially more accurate PMI estimation models. We curated a sample of 213 human subjects from a large-scale human decomposition photographic collection, evaluated accuracy of existing PMI and ADD formulas, and fitted increasingly sophisticated models with additional predictors to estimate the PMI and ADD. Including the total decomposition score (TDS), demographic factors (age, biological sex, and BMI), and weather-related factors (i.e., season of discovery and humidity history) substantially reduced the error of the PMI and ADD models by over 50%. The best performing PMI estimation model used the TDS, demographic, and weather-related features as predictors had an adjusted R-squared of 0.42 and an RMSE of 0.76. It had a 15% lower RMSE than a model using only the TDS to predict the PMI and a 54% lower RMSE than the pre-existing PMI formula. The best ADD estimation model, also using the TDS, demographic, and weather-related features as predictors, resulted in an adjusted R-squared of 0.54 and an RMSE of 0.73. It had a 10% lower RMSE than

the model using only the TDS to predict the ADD and a 55% lower RMSE than the pre-existing ADD formula. This work suggests that the accuracy of existing models can be improved without the need to compute or use hard-to-obtain predictors.

5.2 Introduction

Determining the post-mortem interval (PMI), or time since death, is an important task in human remains cases. An accurate estimation of the PMI can significantly narrow down the list of potential decedents the remains might correspond to, facilitating the eventual identification of the individual. In cases involving homicide, law enforcement professionals can utilize the PMI to eliminate potential suspects and validate witness accounts. Additionally, having knowledge of the PMI aids in establishing the spectrum of natural occurrences and environmental influences that impacted the remains, thereby enabling a more comprehensive analysis.

Human decomposition is a natural and continuous process involving the breakdown of tissues after death that can span from weeks to years, contingent upon a range of biological and environmental factors. Comprehending how these factors influence the rate of human decomposition is pivotal in ascertaining the PMI [33, 16, 57].

Over the years, researchers in the field of forensic anthropology have developed methods that divide the human decomposition process into high-level categories or stages of decay. Each stage is characterized by distinct decomposition phenomena, aiding in the estimation of the PMI [50, 16]. One such method was developed by Gelderman et al. [17], a simplified version of Megyesi et al.’s [33] method, which involves a human decomposition scoring method and regression formulas to estimate the PMI and the accumulated degree days (ADD). The scoring method separates the human body into distinct anatomical regions, which are each assigned a score reflecting the amount of decomposition that has occurred. The individual scores are summed to obtain the total decomposition score (TDS), which is then used to estimate the PMI and ADD using their developed formulae. The ADD is the sum of the average daily ambient temperatures (rather than calendar days) between the date of death and

discovery. Specifically, it represents the heat energy units required for the chemical and biological reactions that decompose a body. Incorporating temperature data and treating decomposition as a semi-continuous variable enabled the measurement of the impact of temperature on decomposition rates. This approach improved PMI estimates by modeling decomposition as dependent on both time and accumulated temperature [33, 17, 52]. In their work, Gelderman et al. [17] developed two PMI estimation formulas: one for directly estimating the PMI and another for estimating the ADD, from which the PMI could later be derived. For simplicity, “PMI/ADD” will be used to refer to both formulae.

Gelderman et al.’s [17] PMI/ADD estimation formulae only considered the TDS as a predictor variable, but what about other factors, such as decedent demographics and environmental information? It is known from existing literature that a wide variety of factors, including body mass [30, 54], temperature [23, 57], humidity [57, 9], and insect activity [52, 46] also affect the decomposition process and should be taken into consideration when determining the PMI. Therefore, expanding the TDS-only PMI/ADD formulae by including such factors could potentially improve their prediction accuracy.

Additionally, Gelderman et al.’s [17] PMI/ADD estimation formulas were developed and evaluated on a very small sample (12 outdoor cases) of human subjects. In fact, it is a crucial factor that not only impacts the study’s statistical power and precision but also influences its capacity to draw meaningful conclusions and apply the results to a wider context [15, 53].

The aim of this study is to investigate ways to improve decomposition-based PMI prediction by (a) using a much larger sample size, (b) employing more flexible and advanced linear models, and (c) enhancing the TDS-only models with demographic and environmental factors known to affect the human decay process. Specifically, various univariate and multivariate linear regression analyses using a combination of predictor variables to estimate the PMI and ADD will be conducted and evaluated. Additionally, Gelderman et al.’s [17] decomposition scoring method will be used

to measure decomposition and calculate the TDS, and their PMI/ADD estimation formulae will be evaluated.

5.3 Materials and methods

5.3.1 The Human Decomposition Photographic Collection

The human decomposition photographic collection is a large-scale image dataset from which the study sample is obtained. The collection includes images of decomposing bodies donated to the Forensic Anthropology Center at The University of Tennessee, Knoxville. The Center houses the Anthropology Research Facility (ARF), an outdoor decomposition laboratory. Forensic experts from the ARF captured these images at non-uniform intervals, with one or more days between each capture. The images, taken from various angles, depict different anatomical areas to illustrate the various stages and regions of human decomposition. The image resolutions vary from 2400×1600 up to 4900×3200 . The dataset covers the period from 2011 to 2023 and comprises over 1.5 million images contributed by more than 800 donors.

5.3.2 The Study Sample

The study sample was obtained from the large-scale human decomposition image dataset (see Section 5.3.1). Specifically, the subjects were chosen with respect to the following attributes:

- biological sex: female or male.
- age groups defined by subtracting the age of the youngest from the age of the oldest donor and dividing by three: $\text{age} < 49$ (younger), $49 \leq \text{age} < 72$ (middle), $\text{age} \geq 72$ (older).

- body mass index (BMI) groups defined by the Center for Disease Control and Prevention (CDC): underweight ($\text{BMI} < 18.5$), healthy ($18.5 \leq \text{BMI} < 25$), overweight ($25 \leq \text{BMI} < 30$), and obese ($\text{BMI} \geq 30$).
- PMI less than 6 months.

By considering the above attributes, the objective was to create a study sample that was both diverse and balanced, meaning it included subjects across both sexes, different age and BMI groups, and PMIs. The final study sample consisted of 213 subjects, with the attribute distribution plots shown in Figure Figure 5.1. Out of the 213 subjects, 108 were female and 105 were male, with ages ranging from 26 to 96 years. All subjects had a known PMI of less than six months.

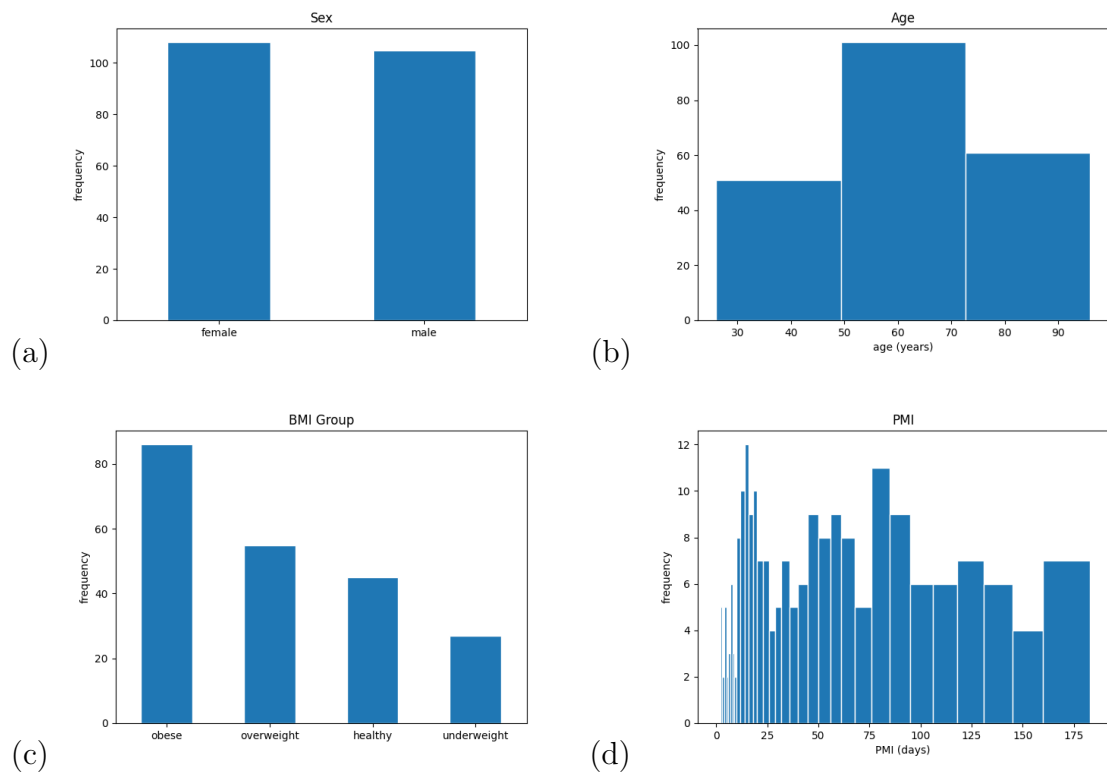


Figure 5.1: Study sample (n=213) distribution plots by (a) sex, (b) age, (c) BMI group, and (d) PMI.

All subjects decomposed outdoors unclothed and intact, meaning no missing body parts so that the decomposition could be scored for the entire body in a standardized fashion. Additionally, no buried, burned, or submerged remains were used in this study because they decompose differently [56]. Since all cases came from the ARF, they decomposed in a similar environmental setting an open-wooded, mostly shaded area with the ground consisting of soil, gravel, and dead/decaying plant matter.

5.3.3 Decomposition Measurement

The state of decomposition was measured for each subject using the human decomposition scoring method proposed by Gelderman et al. [17] (see Appendices B.4 to B.6). To account for the differential decomposition that occurs in different body segments, this scoring method separates the human body into three anatomical regions: (1) the head (including the neck), (2) the torso, and (3) the limbs (including the hands and feet). Based on the morphological features present, each anatomical region is categorized into six stadia or stages, with the lowest (i.e., score 1) indicating no visible changes and the highest (i.e., score 6) indicating complete skeletonization. Once each anatomical region is scored, the three decomposition scores are summed to obtain the total decomposition score (TDS), which ranges from 3 to 18.

The decomposition scoring was performed by author PD, a forensic anthropologist who was well-trained in Gelderman et al.'s [17] scoring method. Specifically, the study sample photos (i.e., photos of the head/neck, torso, and limbs per subject) were assessed and scored on an in-house developed data visualization and annotation software called ICPUTRD (Image Cloud Platform for Use in Tagging and Research on Decomposition) [42]. Once the decomposition scoring was completed, the TDS was calculated for each subject in the study sample.

5.3.4 Weather Data and ADD calculation

Since the ADD in addition to the PMI will be estimated, temperature data was collected from to the closest National Weather Service station at McGhee Tyson Airport, Knoxville, TN. All temperature data was in the form of daily averages and recorded in degrees Celsius (C). Following Gelderman et al. [17], the ADD was calculated for each subject by adding the average daily temperatures above 0 degrees C (base temperature) from death until discovery.

5.3.5 PMI/ADD Estimation

For the PMI/ADD estimation the following were conducted: (1) evaluate Gelderman et al.'s [17] PMI/ADD formulas (5.1) and (5.2) on our larger study sample and (2) perform various univariate and multivariate linear regression analyses using different combinations of the following input variables to predict the PMI and ADD:

- TDS
- Subject demographics, including biological sex, age at death, and BMI.
- Weather-related features, including season of discovery (winter, spring, summer, or fall) and humidity history.

$$\text{PMI} = 10^{(-0.93 + (0.18 \times \text{TDS}))} \quad (5.1)$$

$$\text{ADD} = 10^{(0.03 + (0.19 \times \text{TDS}))} \quad (5.2)$$

In order to be used in the regression analysis, the categorical variables, sex and season of discovery, were converted to numeric by using one-hot-encoding. Specifically, one-hot-encoding creates a dummy variable for each category of the categorical variable with values 1 or 0 representing the presence or absence of the

category. When including dummy variables in a regression analysis one needs to be aware of the the dummy variable trap, which is when two or more variables are highly correlated; in simple terms one variable can be predicted from the others. For instance, knowing male=0, implies female=1, or knowing spring=0, summer=0, fall=0, implies winter=1, and vice versa. The solution to the dummy variable trap is to drop one of the categorical variables, that is, if there are m number of categories, use m-1 in the regression analysis. In doing so, the categorical sex variable was converted to one dummy variable with values 1 for male and 0 for not male (hence female). Similarly, the categorical season of discovery variable was converted to numeric by creating three dummy variables (i.e., spring, summer, and fall), each with values 1 and 0. The BMI was calculated from the estimated cadaver weight (in pounds) and height (in inches). To create the humidity history feature, humidity (relative) data were collected from the nearest National Weather Service station located at McGhee Tyson Airport, Knoxville, TN. Specifically, the humidity history feature was calculated as follows for each subject in the study sample: (1) obtain daily average humidity values for the two weeks prior to the date of discovery (or in this case, when the photograph was taken) and (2) calculate the average for the two-week humidity history, resulting in the numeric variable called hum_hist. The two-week humidity history was aggregated to produce a single summary value (i.e., the average humidity) to be used in the regression model. Future work should further evaluate this approach by using different humidity history ranges (e.g., three weeks, four weeks, etc.) and various aggregation methods (e.g., median, mode, standard deviation, etc.). Also, a two-week temperature history (i.e., temp_hist) was initially included, but it was dropped from the regression models due to multicollinearity with other weather-based variables.

The statistical analysis was conducted using SPSS Statistics (version 29.0.2.0) and the Python package, statsmodels. Specifically the following univariate and multivariate linear regression analysis using the study sample of 213 subjects were conducted and compared:

1. TDS to predict the PMI and ADD (univariate).
2. TDS + demographic features (sex, age, and BMI) to predict the PMI and ADD (multivariate).
3. TDS + demographic features (sex, age, and BMI) + weather features (spring, summer, fall, and hum_hist) to predict the PMI and ADD (multivariate).

Similar to Gelderman et al. [17], for the linear regression analysis, the PMI and ADD were naturally log-transformed to achieve a more linear relationship between the TDS and PMI, and the TDS and ADD as shown by Figures 5.2 to 5.5.

Additionally, all continuous independent variables, including the TDS, age, BMI, and hum_hist, were standardized to be on the same scale and thus contribute equally to the regression analysis. Z-score standardization was applied, rescaling the original variable to have a mean of zero and a standard deviation of one. Mathematically, this involves subtracting the mean of the original variable from the raw value and then dividing it by the standard deviation of the original variable, as shown in equation (5.3).

$$\text{standardized } X = \frac{X - \text{mean}}{\text{standard deviation}} \quad (5.3)$$

The metrics used to evaluate the different regression analyses were the adjusted R-squared (adj. R^2) and root mean squared error (RMSE). The adj. R^2 , as shown by equation (5.5), is comparable to the R^2 or the coefficient of determination, as shown by equation (5.4), as its value lies between 0 and 1 (the higher, the better) and explains the proportion of variance for a dependent variable with respect to an independent variable(s) in the regression model.

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} \quad (5.4)$$

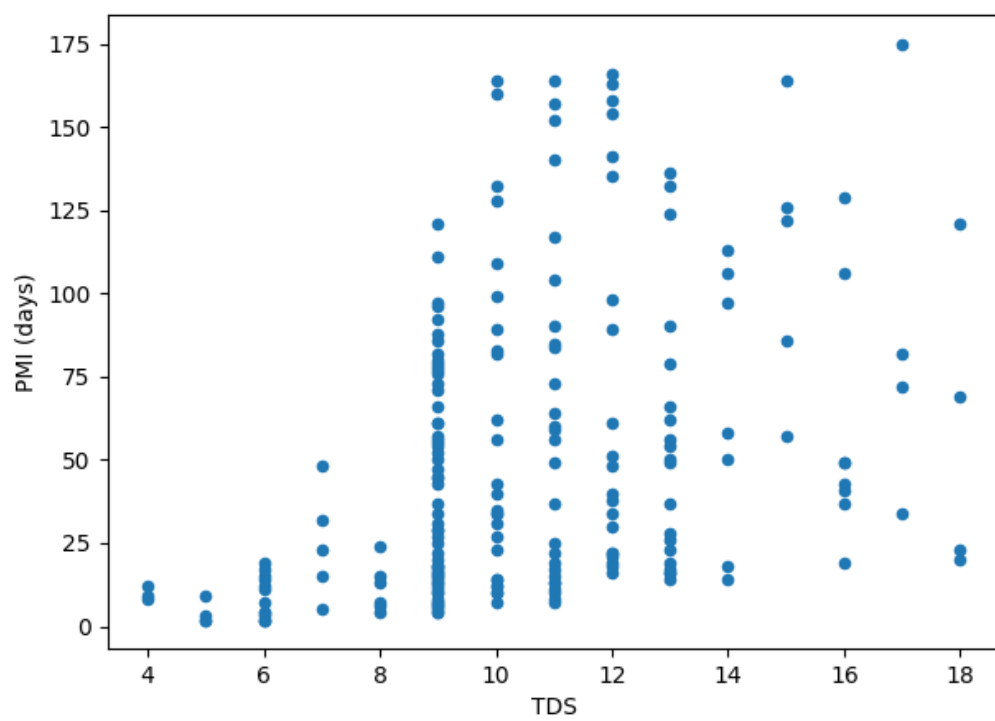


Figure 5.2: Plot of TDS vs. PMI (n=213).

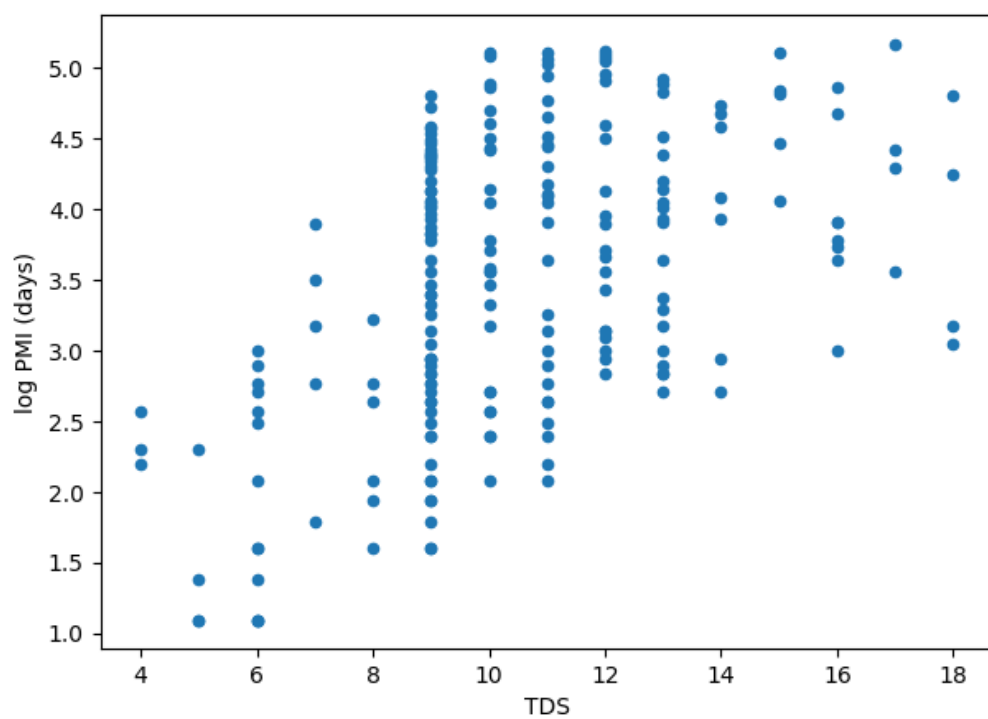


Figure 5.3: Plot of TDS vs. naturally log-transformed PMI (n=213).

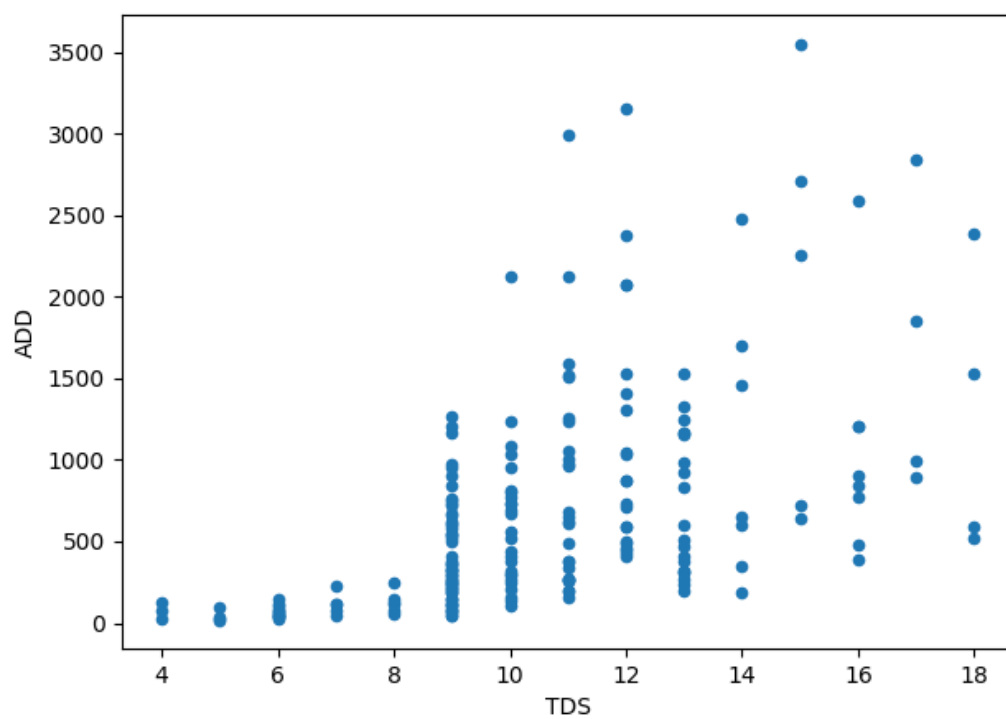


Figure 5.4: Plot of TDS vs. ADD (n=213).

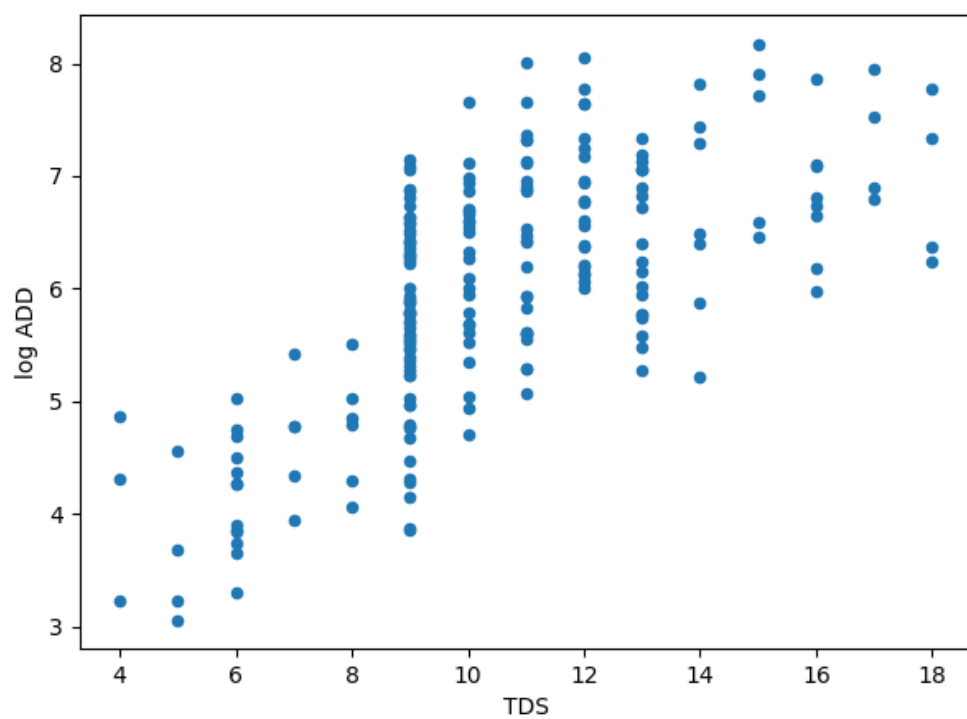


Figure 5.5: Plot of TDS vs. naturally log-transformed ADD (n=213).

Where:

RSS = residual sum of squares

TSS = total sum of squares

y_i = actual value of sample i

$\hat{y}_i$ = predicted value of sample i

$\bar{y}$ = mean value of y

$$adj.R^2 = \frac{(1 - R^2)(n - 1)}{(n - k - 1)} \quad (5.5)$$

Where:

n = number of samples

k = number of independent variables

The difference between R^2 and adj. R^2 is that R^2 assumes all independent variables considered affect the model's outcome, whereas adj. R^2 accounts only for those variables that actually impact the model's performance. Adj. R^2 penalizes the model for including unnecessary variables that do not significantly improve its accuracy. For instance, when multiple linear regression models are built, such as in this study with the forward addition method, at each iteration independent variables are added, the R^2 will keep increasing, but the adj. R^2 will only increase when the variable actually affects the dependent variable. If a variable is non-significant, the R^2 value will still increase, but the adj. R^2 value will decrease at that point. The RMSE, as shown by equation (6), is the standard deviation of the residuals, measuring the average difference between a model's predicted values and the actual values.

$$RMSE = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n}} \quad (5.6)$$

Where:

n = number of samples

y_i = actual value of sample i

$\hat{y}_i$ = predicted value of sample i

Residuals represent the distance between the regression line and the data points. As the data points move closer to the regression line, the model has less error, lowering the RMSE and producing more precise predictions. The RMSE values range from 0 (perfect predictions) to positive infinity and are in the same units as the dependent variable. Note, the RMSE is almost identical to the standard error of the regression. The difference is that the standard error adjusts for the number of independent variables while RMSE does not. The regression coefficients and their p-values (with a significance level of .05) were calculated and reported. Since the predictor variables were standardized, the regression coefficients can be interpreted as the standardized regression coefficients. Finally, predicted vs. actual scatter plots will be created to visualize and compare the fit of the different regression analyses. In these plots, the x-axis represents the actual values, and the y-axis represents the predicted values. Ideally, if the predictions are perfect, the points will lie along a straight line with a slope of 1.

5.4 Results

Table 5.1 shows the PMI/ADD estimation results.

Table 5.1: The univariate and multivariate linear regression analysis results. Reported are the adj. R^2 , the RMSE, and the back-transformed (i.e., from the log scale back to the normal scale) formula with the standardized regression coefficients.

Analysis (predictors, output)	adj. R^2	RMSE	Formula (back-transformed)
TDS, PMI (Gelderman et al. [17])	-	1.67	See (5.1)
TDS, ADD (Gelderman et al. [17])	-	1.62	See (5.2)
TDS, PMI	.23	.89	$PMI = \exp(1.68 + .17TDS)$
TDS, ADD	.45	.81	$ADD = \exp(3.27 + .26TDS)$
TDS+demographics, PMI	.25	.87	$PMI = \exp(3.58 + .5TDS + .14BMI + .02age - .17sex)$
TDS+demographics, ADD	.47	.78	$ADD = \exp(5.97 + .73TDS + .19BMI + .03age + .04sex)$
TDS+demographics+weather, PMI	.42	.76	$PMI = \exp(3.79 + .63TDS + .12BMI - .01age - .07sex + .1spring - .55summer - .77fall - .15hum_hist)$
TDS+demographics+weather, ADD	.54	.73	$ADD = \exp(5.58 + .68TDS + .17BMI + .02age + .06sex + .54spring + .62summer + .16fall - .12hum_hist)$

The first two rows (above the dashed line) give the results for Gelderman et al.'s [17] PMI and ADD formulas evaluated on our study sample. To ensure equal comparison among all regression experiments, the log-transformed versions of Gelderman et al.'s [17] PMI and ADD formulas were used, which required applying the base 10 logarithm to both sides of (5.1) and (5.2). This resulted in an RMSE of 1.67 when estimating the PMI and an RMSE of 1.62 when estimating the ADD. Note, the adj. R^2 was not calculated here since these formulae were evaluated and not fitted on our study sample. The remaining rows of Table 5.1 give the results for the univariate and multivariate linear regression using different sets of input variables. For the univariate linear regression analysis using only the TDS as a predictor variable of PMI, an adj. R^2 of .23 and a RMSE of .89, and for ADD, an adj. R^2 of .45 and a RMSE of .81, were achieved. In both cases, the standardized coefficients were statistically significant with a p-value $< .05$. For the multivariate linear regression analysis using the TDS and demographic features as predictor variables of PMI, an adj. R^2 of .25 and a RMSE of .87, and for ADD, an adj. R^2 of .47 and a RMSE of .78, were achieved. In both cases, the standardized coefficients for BMI had a p-value $< .05$, hence statistically significant, and a p-value $> .05$ for sex and age, hence not statistically significant. Lastly, for the multivariate linear regression analysis using the TDS, demographic, and weather features as predictor variables of PMI, an adj. R^2 of .42 and a RMSE of .76, and for ADD, an adj. R^2 of .54 and a RMSE of .73, were achieved. In both cases, the standardized coefficients of the weather features (i.e., spring, summer, fall, and hum_hist) and BMI had a p-value $< .05$, hence statistically significant, and a p-value $> .05$ for the sex and age, hence not statistically significant.

Figure 5.6 shows the predicted vs. actual PMI scatter plots for the regression analyses with input variables: (a) TDS, (b) TDS + demographic features, (c) TDS + demographic features + weather features.

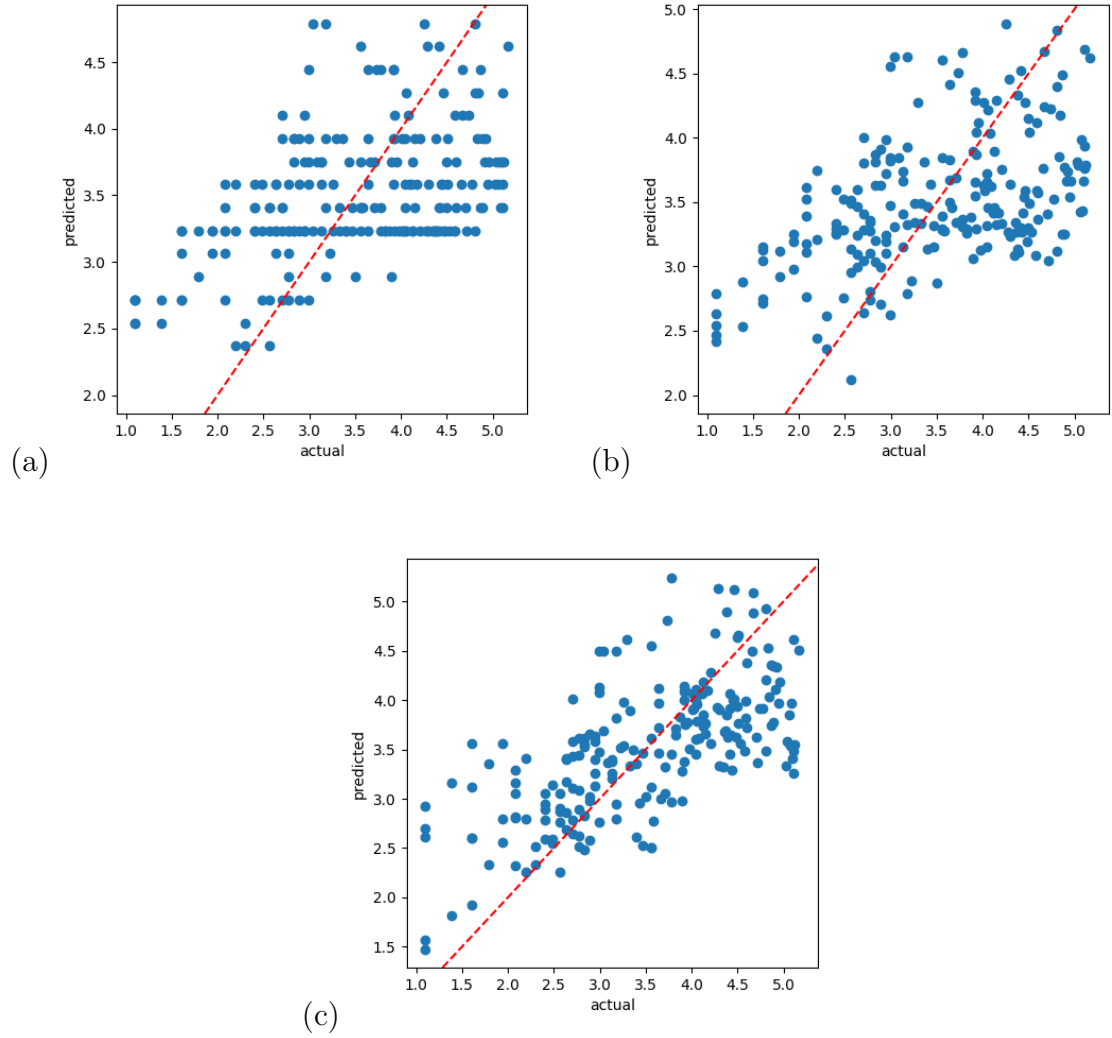


Figure 5.6: Predicted vs. actual PMI scatter plots for the different regression analyses with input variables: (a) TDS, (b) TDS + demographic features, (c) TDS + demographic features + weather features.

Similarly, Figure 5.7 shows the predicted vs. actual ADD scatter plots for the regression analyses with input variables with input variables: (a) TDS, (b) TDS + demographic features, (c) TDS + demographic features + weather features. The red dashed line gives the reference line $x=y$ (i.e., $\text{actual}=\text{predicted}$) with slope 1.

5.5 Discussion

The objective of this study was to investigate ways to improve decomposition-based PMI estimation from human remains by (a) using a much larger sample size, (b) employing more flexible and advanced linear models, and (c) enhancing PMI/ADD estimation models with additional factors, such as demographic and weather-related data. Specifically, pre-existing PMI/ADD formulas proposed by Gelderman et al. [17] were evaluated on the significantly larger sample size. Additionally, various univariate and multivariate linear regression analyses using a variety of predictor variables to estimate the PMI and ADD were conducted and evaluated.

When evaluating Gelderman et al.'s [17] outdoor PMI/ADD formulas on our study sample, the results were poor. This is not surprising given that the formulae were developed from a very small sample size (i.e., 12 outdoor cases compared to the 213 outdoor cases of this study). In fact, Gelderman et al.'s [17] PMI formula resulted in a RMSE that was 54% higher than the RMSE of the PMI model that considered the TDS, demographic, and weather-related features as predictor variables. Similarly, Gelderman et al.'s [17] ADD formula resulted in a RMSE that was 55% higher than the RMSE of the ADD model that considered the TDS, demographic, and weather-related features as predictor variables.

The univariate and multivariate linear regression analysis results showed that including additional factors, such as demographic and weather-related information, in addition to the TDS, improves PMI/ADD prediction performance. When estimating the PMI, the adjusted R^2 was 45% higher and the RMSE was 15% lower when using the TDS, demographic, and weather-related features compared to only using the TDS

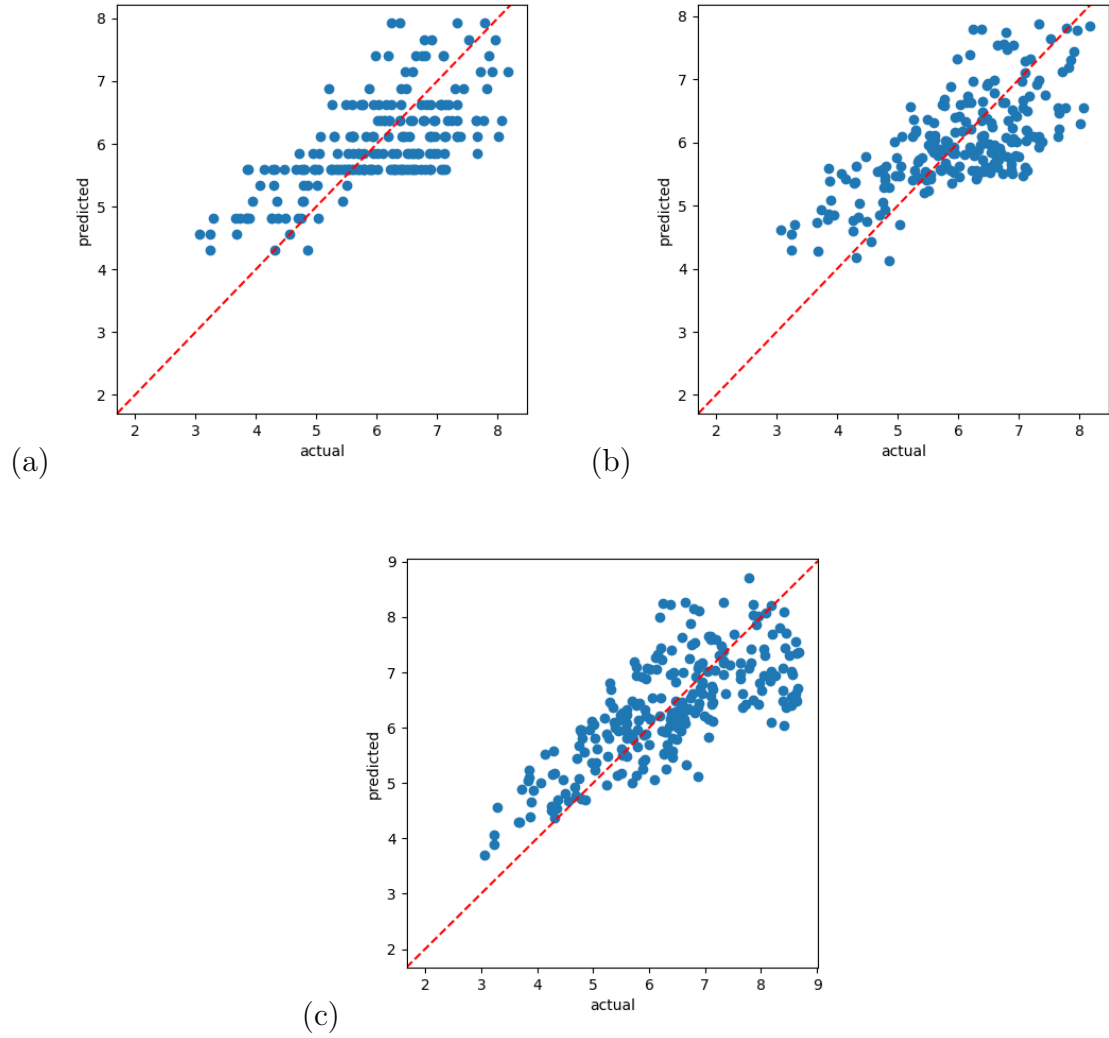


Figure 5.7: Predicted vs. actual ADD scatter plots for the different regression analyses with input variables: (a) TDS, (b) TDS + demographic features, (c) TDS + demographic features + weather features.

as predictor variables. Similarly, when estimating the ADD, the adjusted R^2 was 17% higher and the RMSE was 10% lower when using the TDS, demographic, and weather-related features compared to only using the TDS as predictor variables. This increase in prediction performance is also seen in the predicted vs. actual scatter plots (see Figures 5.6 and 5.7) as more predictor variables are considered, the points are closer to the red-dashed line (reference line), indicating a better fit and more accurate predictions. When assessing variable importance, the standardized coefficient of the TDS was statistically significant for the TDS-only PMI/ADD models. When considering the demographic features (i.e., age, sex, and BMI) in addition to the TDS, the standardized coefficients for TDS and BMI were statistically significant, but not for the age and sex variables when predicting the PMI and ADD. Lastly, when also considering the weather-related features (i.e., season of discovery and humidity history), the TDS, BMI, and all weather-related features had statistically significant standardized coefficients, but not the age and sex. This suggests that the demographic features, sex and age, may not have a significant effect on the dependent variables, PMI and ADD, and could thus be removed from the model.

While the methods described in this research build upon and provide improvements to previous methods, there are some important limitations of this work. For one, the decomposition measurement to obtain the TBS for each subject in the study sample was conducted by a single forensic expert. This may introduce labeling bias and errors, in particular for large sample sizes, resulting in less accurate models [22, 40]. Therefore, future work will focus on creating a so-called “gold standard” TDS dataset—a dataset carefully labeled and evaluated by multiple forensic experts. Such a dataset would be accepted as the most accurate and reliable of its kind. This step will ensure that the models are created with accurate and unbiased data, which is vital for developing high-quality models.

An additional limitation is that this study is environmental- and climate-specific. All study sample subjects decayed outdoors in an open-wooded area with the ground consisting of soil, gravel, and dead/decaying plant matter (e.g., rotting wood

and shedding leaves). Additionally, the climate of this area is humid subtropical, characterized by high summer and moderate winter temperatures. Therefore, the PMI/ADD formulas of this study may not perform as well for different climate conditions or environments.

And finally, the PMI/ADD formulas reported in this study should be used with caution. The aim of this work was not to create new PMI/ADD estimation formulas but to investigate ways to improve decomposition-based PMI estimation methods, such as the one proposed by Gelderman et al. [17], by using a much larger sample size and including different factors known to affect the human decay process. Researchers are encouraged to evaluate and build upon our method to further assess its accuracy. Other existing decomposition-based PMI estimation formulae, such as those developed by Megyesi et al. [33] and Moffat et al. [34], could also be modified and evaluated using additional variables as proposed by this work.

5.6 Conclusion

In conclusion, this study underscores the significance of leveraging larger datasets and incorporating a broader range of predictive factors to enhance PMI/ADD estimation models in forensic science. By integrating demographic and environmental variables with the TDS, the accuracy of PMI/ADD predictions compared to traditional models improved. The results suggest that a more comprehensive approach, which accounts for factors such as age, biological sex, BMI, season, temperature, and humidity, can better capture the complexity of human decomposition. This advancement not only highlights the potential for more accurate post-mortem interval assessments but also lays the groundwork for future research to further refine and validate these models.

Chapter 6

Conclusion

6.1 Research Overview

The current research aimed to make a human decomposition photographic collection accessible and useful for forensic research and casework, as well as to develop AI tools focused on post-mortem interval estimation. The contributions of my dissertation can be summarized as follows:

1. Continue to develop and maintain a collaborative tool for forensic research and casework that facilitates the organization, visualization, analysis, and curation of a large collection of human decomposition photographs.
2. Develop a method to perform automatic stage of decay labeling when labeled data is limited, and subsequently, develop AI models to automate stage of decay identification.
3. Evaluate existing PMI estimation methods using our human decomposition data set, which allows for a much larger sample size than earlier studies.

6.2 Discussion, Contributions, and Future Work

This section discusses the key findings and future directions of the four core chapters(2,3,4,5). Chapter 2 focused on the continued development and evaluation of a collaborative web-based software system, called ICPUTRD, for data organization, analysis, and annotation to support this and other research studies in forensic science and casework. The study results confirmed the feasibility and practicality of ICPUTRD in facilitating forensic research and casework involving a large-scale photographic collection documenting human decomposition. ICPUTRD is currently installed and operational at the Forensic Anthropology Department and the Anthropology Research Facility at the University of Tennessee, Knoxville. Future work will focus on maintaining and further developing ICPUTRD to adapt to the needs of upcoming projects and users.

In Chapter 3, the aim was to train models for predicting the stage of human decay while addressing the challenge of limited labeled data. As a solution, a supervised vision system was developed, incorporating a domain-aware label propagation algorithm along with various state-of-the-art CNN classification models. The experiments demonstrated that using the developed label propagation algorithm resulted in the highest performance in stage-of-decay classification. This method has been employed to automatically label additional images from the human decomposition photographic collection, significantly reducing manual labeling efforts and associated costs.

Chapter 4 extended the research from Chapter 3 by exploring the feasibility of automating two established human decomposition scoring methods using AI. Additionally, an interrater study was conducted to assess the reliability of the developed AI models compared to human forensic examiners. The results demonstrated the potential of AI models, trained on the human decomposition image dataset, to automate the identification of stages of decay with reliability comparable to that of a human expert. While the results were promising, there are some important limitations to this work. Firstly, the data labeling was performed by a single forensic expert, which may introduce labeling bias and lead to biased training datasets. Future research should focus on addressing this limitation to ensure that the models are trained with accurate and unbiased data, which is vital for developing high-quality models. Another limitation is that this study is environment-specific, as the AI models were trained using images of donors decaying outdoors in an open-wooded area. As a result, the AI models may not perform as well on images taken in different environments, requiring additional training and retraining. The models developed in this chapter have been uploaded and deployed to the project's Hugging Face organization*, contributing both to our database of forensic AI tools and to the open-source data science and machine learning community.

*<https://huggingface.co/icputrd>

Once the stage of decay is identified, this information is used to help estimate the PMI. Decomposition-based PMI estimation methods exist, but they were either developed and/or evaluated on relatively small sample sizes, often resulting in very low accuracy. Additionally, these methods were limited from a data perspective, as they only considered a single predictor variable, despite literature showing that multiple factors, including donor demographics and environmental information, can affect the human decay process. Therefore, the final core chapter, Chapter 5, focused on (1) evaluating an existing PMI estimation method using the significantly larger sample size available to us and (2) investigating ways to improve these methods by incorporating additional data known to affect human decomposition, thereby employing more flexible and advanced models. The study results underscored the significance of leveraging larger datasets and incorporating a broader range of predictive factors to better capture the complexity of human decomposition.

In the future, I believe the ever-challenging task of PMI estimation should be approached with a non-linear solution. For instance, the use of large language models (LLMs), which have gained significant popularity in recent years, presents an exciting avenue for PMI estimation by leveraging multimodal data. By integrating visual data from images of remains with demographic, decomposition, and environmental data, LLMs can be trained to recognize complex patterns in decomposition that a simple linear regression model cannot capture. While linear regression relies on linear relationships between variables, LLMs can identify intricate patterns in image data and the interactions among multiple features, which are not necessarily linear. Additionally, LLMs can leverage vast amounts of unstructured data (e.g., images and text), learning from it without requiring extensive manual feature engineering, unlike linear regression.

6.3 Conclusion

In conclusion, this research contributes to the growing body of forensic science by enhancing the utility of a vast photographic database of human decomposition for both research and casework. By developing tools for organizing, visualizing, and analyzing decomposition data, this work supports more efficient forensic investigations. The development of AI methods for stage of decay labeling and PMI estimation opens new possibilities for automating forensic processes and improving accuracy in PMI predictions. Ultimately, this research lays the foundation for more reliable and scalable forensic tools, advancing the integration of AI in forensic science.

Bibliography

- [1] Angular. <https://angular.io>. Accessed: 2024-03-20. 15
- [2] Annotorious. <https://recogito.github.io/annotorious/>. Accessed: 2024-03-20. 17
- [3] S. Balaji and M. S. Murugaiyan. Waterfall vs. v-model vs. agile: A comparative study on sdlc. *International Journal of Information Technology and Business Management*, 2(1):26–30, 2012. 12
- [4] N. Bevan and M. Macleod. Usability measurement in context. *Behaviour & information technology*, 13(1-2):132–145, 1994. 18
- [5] J. Brooke. Sus: a “quick and dirty” usability scale. In *Usability evaluation in industry*, pages 107–115. Taylor & Francis, London, 1996. 18
- [6] F. Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017. 41, 52
- [7] R. Ciaffi, A. Feola, E. Perfetti, S. Manciocchi, S. Potenza, G. L. Marella, et al. Overview on the estimation of post mortem interval in forensic anthropology: review of the literature and practical experience. *Romanian Journal of Legal Medicine*, 26(4):403–411, 2018. 10
- [8] A. Cockburn. *Agile software development: the cooperative game*. Pearson Education, London, second edition, 2006. 12
- [9] D. L. Cockle and L. S. Bell. Human decomposition and the reliability of a ‘universal’ model for post mortem interval estimations. *Forensic science international*, 253:136:e1–136:e9, 2015. 66
- [10] G. S. Cooper and V. Meterko. Cognitive bias research in forensic science: a systematic review. *Forensic science international*, 297:35–46, 2019. 3, 48

- [11] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE conference on computer vision and pattern recognition*, pages 248–255, 2009. [41](#), [52](#)
- [12] A. Deshpande and M. Kumar. *Artificial intelligence for big data: complete guide to automating big data solutions using artificial intelligence techniques*. Packt Publishing Ltd, 2018. [48](#)
- [13] P. D. Ditto. Human decomposition evaluation: A standardized approach for staging and scoring morphological features using artificial intelligence. 2024. [48](#)
- [14] Express. <https://expressjs.com/>. Accessed: 2024-03-20. [15](#)
- [15] J. Faber and L. M. Fonseca. How sample size influences research outcomes. *Dental press journal of orthodontics*, 19:27–29, 2014. [66](#)
- [16] A. Galloway, W. H. Birkby, A. M. Jones, T. E. Henry, and B. O. Parks. Decay rates of human remains in an arid environment. *Journal of Forensic Science*, 34(3):607–616, 1989. [3](#), [36](#), [48](#), [65](#)
- [17] H. Gelderman, L. Boer, T. Naujocks, A. IJzermans, and W. Duijst. The development of a post-mortem interval estimation for human remains found on land in the netherlands. *International journal of legal medicine*, 132(3):863–873, 2018. [ix](#), [xii](#), [3](#), [4](#), [33](#), [48](#), [49](#), [51](#), [52](#), [53](#), [55](#), [56](#), [57](#), [58](#), [59](#), [60](#), [61](#), [65](#), [66](#), [70](#), [71](#), [73](#), [80](#), [81](#), [83](#), [86](#), [106](#), [107](#), [108](#), [109](#)
- [18] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. [41](#)
- [19] M. Huo, J. Verner, L. Zhu, and M. A. Babar. Software quality and agile methods. In *Proceedings of the 28th Annual International Computer Software and Applications Conference, 2004. COMPSAC 2004.*, pages 520–525. IEEE, 2004. [12](#)

- [20] Imagehash. <https://github.com/JohannesBuchner/imagehash>. Accessed: 2024-09-01. [xviii](#), [102](#)
- [21] M. H. Jarrahi. Artificial intelligence and the future of work: human-ai symbiosis in organizational decision making. *Business horizons*, 61(4):577–586, 2018. [48](#)
- [22] H. Jiang and O. Nachum. Identifying and correcting label bias in machine learning. In *International conference on artificial intelligence and statistics*, pages 702–712. PMLR, 2020. [61](#), [85](#)
- [23] A. P. Johnson, K. M. Mikac, and J. F. Wallman. Thermogenesis in decomposing carcasses. *Forensic science international*, 231(1-3):271–277, 2013. [66](#)
- [24] J. H. Korteling, G. C. van de Boer-Visschedijk, R. A. Blankendaal, R. C. Boonekamp, and A. R. Eikelboom. Human-versus artificial intelligence. *Frontiers in artificial intelligence*, 4:1–13, 2021. [48](#)
- [25] J. Kukucka, S. M. Kassin, P. A. Zapf, and I. E. Dror. Cognitive bias and blindness: a global survey of forensic science examiners. *Journal of applied research in memory and cognition*, 6(4):452–459, 2017. [3](#), [48](#)
- [26] J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, pages 159–174, 1977. [xi](#), [xii](#), [56](#), [58](#), [59](#)
- [27] M. Lin, Q. Chen, and S. Yan. Network in network. *arXiv preprint arXiv:1312.4400*, 2013. [53](#)
- [28] R. W. Mann, W. M. Bass, and L. Meadows. Time since death and decomposition of the human body: variables and observations in case and experimental field studies. *Journal of forensic sciences*, 35(1):103–111, 1990. [4](#)
- [29] S. J. Marhoff, P. Fahey, S. L. Forbes, and H. Green. Estimating post-mortem interval using accumulated degree-days and a degree of decomposition index in

- australia: a validation study. *Australian Journal of Forensic Sciences*, 48(1):24–36, 2016. [11](#)
- [30] S. Matuszewski, S. Konwerski, K. Fratzczak, and M. Szafałowicz. Effect of body mass and clothing on decomposition of pig carcasses. *International journal of legal medicine*, 128:1039–1048, 2014. [66](#)
- [31] Q. McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947. [43](#)
- [32] Mean.js. <http://meanjs.org>. Accessed: 2024-03-20. [15](#)
- [33] M. S. Megyesi, S. P. Nawrocki, and N. H. Haskell. Using accumulated degree-days to estimate the postmortem interval from decomposed human remains. *Journal of Forensic Science*, 50(3):618–626, 2005. [ix](#), [xi](#), [xii](#), [3](#), [4](#), [11](#), [33](#), [48](#), [49](#), [51](#), [52](#), [53](#), [55](#), [56](#), [57](#), [58](#), [59](#), [60](#), [61](#), [65](#), [66](#), [86](#), [103](#), [104](#), [105](#)
- [34] C. Moffatt, T. Simmons, and J. Lynch-Aird. An improved equation for tbs and add: establishing a reliable postmortem interval framework for casework and experimental studies. *Journal of Forensic Sciences*, 61:S201–S207, 2016. [86](#)
- [35] MongoDB. <https://www.mongodb.com/>. Accessed: 2024-03-20. [15](#)
- [36] S. Mousavi, D. Lee, T. Griffin, D. Steadman, and A. Mockus. Collaborative learning of semi-supervised clustering and classification for labeling uncured data. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 1716–1720. IEEE, 2020. [xviii](#), [102](#)
- [37] J. Myburgh, E. N. L’Abbé, M. Steyn, and P. J. Becker. Estimating the postmortem interval (pmi) using accumulated degree-days (add) in a temperate region of south africa. *Forensic science international*, 229(1-3):165–e1, 2013. [4](#)

- [38] S. Nakhaeizadeh, I. E. Dror, and R. M. Morgan. Cognitive bias in forensic anthropology: visual assessment of skeletal remains is susceptible to confirmation bias. *Science & Justice*, 54(3):208–214, 2014. [3](#), [48](#)
- [39] A.-M. Nau, P. Ditto, D. W. Steadman, and A. Mockus. Identifying factors to help improve existing decomposition-based pmi estimation methods. *arXiv preprint arXiv:2409.09056*, 2024. [64](#)
- [40] A.-M. Nau, P. Ditto, D. W. Steadman, and A. Mockus. Towards automation of human stage of decay identification: An artificial intelligence approach. *arXiv preprint arXiv:2408.10414*, 2024. [47](#), [85](#)
- [41] A.-M. Nau, A. Mockus, and D. W. Steadman. Stage of decay estimation exploiting exogenous and endogenous image attributes to minimize manual labeling efforts and maximize classification performance. In *IEEE international conference on image processing*, pages 2705–2709. IEEE, 2023. [60](#)
- [42] A.-M. Nau, S. Mousavi, D. Lee, R. Hossain, T. Griffin, D. W. Steadman, and A. Mockus. Icpurtd: image cloud platform for use in tagging and research on decomposition. *Journal of forensic sciences*, 2023. [52](#), [56](#), [70](#)
- [43] J. Nielsen, R. L. Mack, and editors. *Usability inspection methods*. Pearson Education, New York, NY, 1994. [18](#)
- [44] Node.js. <https://nodejs.org/en/>. Accessed: 2024-03-20. [15](#)
- [45] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2010. [41](#), [52](#)
- [46] J. A. Payne. A summer carrion study of the baby pig *sus scrofa linnaeus*. *Ecology*, 46(5):592–602, 1965. [66](#)
- [47] A.-I. Piraianu, A. Fulga, C. L. Musat, O.-R. Ciobotaru, D. G. Poalelungi, E. Stamate, O. Ciobotaru, and I. Fulga. Enhancing the evidence with algorithms:

- how artificial intelligence is transforming forensic medicine. *Diagnostics*, 13(18):2992–3003, 2023. [4](#), [48](#)
- [48] U. N. Raghavan, R. Albert, and S. Kumara. Near linear time algorithm to detect community structures in large-scale networks. *Physical review E*, 76(3):036106, 2007. [xi](#), [40](#), [41](#), [42](#), [43](#), [44](#)
- [49] A. M. Rahmani, E. Azhir, S. Ali, M. Mohammadi, O. H. Ahmed, M. Y. Ghafour, S. H. Ahmed, and M. Hosseinzadeh. Artificial intelligence approaches and mechanisms for big data analytics: a systematic study. *PeerJ computer science*, 7:1–28, 2021. [48](#)
- [50] W. C. Rodriguez and W. M. Bass. Insect activity and its relationship to decay rates of human cadavers in east tennessee. *Journal of forensic sciences*, 28(2):423–432, 1983. [3](#), [65](#)
- [51] K. A. Sauerwein. *Perceptions and cognitive bias in decomposition scoring methods in forensic anthropology*. PhD thesis, University of Tennessee, 2018. [48](#)
- [52] T. Simmons, R. E. Adlam, and C. Moffatt. Debugging decomposition data—comparative taphonomic studies and the influence of insects and carcass size on decomposition rate. *Journal of forensic sciences*, 55(1):8–13, 2010. [66](#)
- [53] L. M. Sullivan, J. Weinberg, and J. F. Keaney Jr. Common statistical pitfalls in basic science research. *Journal of the American Heart Association*, 5(10):e004142, 2016. [66](#)
- [54] A. Sutherland, J. Myburgh, M. Steyn, and P. J. Becker. The effect of body size on the rate of decomposition in a temperate region of south africa. *Forensic science international*, 231(1-3):257–262, 2013. [66](#)
- [55] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. [41](#), [52](#)

- [56] M. A. van Daalen, D. S. de Kat, B. F. Oude Grotebevelsborg, R. De Leeuwe, J. Warnaar, R. J. Oostra, and W. L. M Duijst-Heesters. An aquatic decomposition scoring method to potentially predict the postmortem submersion interval of bodies recovered from the north sea. *Journal of forensic sciences*, 62(2):369–373, 2017. [70](#)
- [57] A. A. Vass. The elusive universal post-mortem interval formula. *Forensic science international*, 204(1-3):34–40, 2011. [3](#), [48](#), [65](#), [66](#)
- [58] D. J. Wescott. Recent advances in forensic anthropology: decomposition research. *Forensic sciences research*, 3(4):278–293, 2018. [10](#)

Appendices

A Chapter 2 Supporting Information

A.1 The Complete Nomenclature.

List of keywords				
adipocere	brown discoloration	glossy skin	mesh bag	saturated soil
advanced decomposition	bruising	gray discoloration	moist decomposition	scale
amputation	bullae	green gut	moist skin	scar
ant	bullae	groin	mold	scatter
ants	cage	hair	mottling	scavenging
arm	clothing	hand	mouse	shoe cover
autopsy bag	darkened skin	head	mouth	skeletonization
autopsy cut	dentures	hernia	mummification	skin slippage
axilla	dermis	hips	mummified	skull
back	desiccated skin	insect	neck	snail
background	desiccation	larvae	nose	soft tissue
backside	discoloration	larval mass	open sore	soil stain
bandage	dry skin	leaf	organ bag	stake
bee	duct tape	left foot	personal effects	stink bug
beetle	duplicate	left hand	post bloat	stomata
beginning bloat	ear	leg	prone	supine
black brown discoloration	eggs	legs	prosthetic device	surgical implant
black discoloration	exposed bone	lividity	purge	surgical item
black gray discoloration	eye	livor mortis	raccoon print	tag
black plastic	face	maggot	recessed eye	tattoo
black tarp	flies	maggot eggs	recognizable face	torso
bloat	fly	maggot foam	remove	trauma
body	foam	maggot mass	remove?	trunk
bone	foot	maggots	right foot	waterlogged skin
bone exposure	fresh skin	marbling	right hand	whole body
brick	genitals	medical device	sampling site	worm
zip tie	yellow jacket			

Figure A.1: The complete nomenclature consisting of 132 forensic-related keywords.

A.2 ETL Data Pipeline.

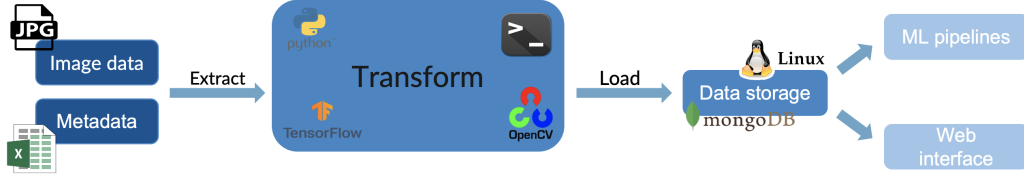


Figure A.2: The ETL data pipeline illustrates ICPUTRD’s data management process. The first step was to extract the raw data, which included all donor images and image metadata, such as donor demographics and temperature and humidity recordings from Tinytag data loggers. Once the raw data was extracted, it was processed, cleaned, and transformed to fit the required structure. Specifically, this involved standardizing filenames, anonymizing donor IDs (i.e., assigning a unique 3-digit hexadecimal number), removing duplicate images using image hashing [20], correcting image orientation, and detecting and removing images containing irrelevant content (e.g., stakes displaying unanonymized donor IDs or plastic bags/cages covering the donor) using a trained CNN classifier to automate this task [36]. Finally, the transformed data was loaded into the target system, involving local storage as well as storage in a MongoDB database.

B Chapter 4 Supporting Information

B.1 Megyesi et al.’s [33] decomposition scoring method for the head.

Table B.1: Megyesi et al.’s [33] stages and description for the head (including the neck).

Stage	Description
Fresh	Fresh, no discoloration
Early decomposition	1. Pink-white appearance with skin slippage and some hair loss.
	2. Gray to green discoloration: some flesh still relatively fresh.
	3. Discoloration and/or brownish shades particularly at edges, drying of nose, ears, lips.
	4. Purging of decompositional fluids out of eyes, ears, nose, mouth, some bloating may be present.
	5. Brown to black discoloration of flesh.
Advanced decomposition	1. Caving in of flesh and tissues of eyes and throat.
	2. Moist decomposition with bone exposure less than one half that of the area being scored.
	3. Mummification with bone exposure less one half that of the area being scored.
Skeletonization	1. Bone exposure of more than half the area scored with greasy substances and decomposed tissue.
	2. Bone exposure of more than half the area scored with desiccated or mummified tissue.
	3. Bones largely dry, but retaining some grease.
	4. Dry bone.

B.2 Megyesi et al.'s [33] decomposition scoring method for the torso.

Table B.2: Megyesi et al.'s [33] stages and description for the torso.

Stage	Description
Fresh	Fresh, no discoloration
Early decomposition	1. Pink-white appearance with skin slippage and marbling present. 2. Gray to green discoloration: some flesh still relatively fresh. 3. Bloating with green discoloration and purging of decomposition fluids. 4. Postbloating following release of abdominal gasses, with discoloration changing from green to black.
Advanced decomposition	1. Decomposition of tissue producing sagging of flesh; caving in of the abdominal cavity. 2. Moist decomposition with bone exposure less than one half that of the area being scored. 3. Mummification with bone exposure less one half that of the area being scored.
Skeletonization	1. Bone with decomposed tissue, sometimes with body fluids and grease still present. 2. Bones with desiccated or mummified tissue covering less than one half of the area being scored. 3. Bones largely dry, but retaining some grease. 4. Dry bone.

B.3 Megyesi et al.'s [33] decomposition scoring method for the limbs.

Table B.3: Megyesi et al.'s [33] stages and description for the limbs.

Stage	Description
Fresh	Fresh, no discoloration
Early decomposition	1. Pink-white appearance with skin slippage of hands and/or feet. 2. Gray to green discoloration: some flesh still relatively fresh. 3. Discoloration and/or brownish shades particularly at edges, drying of fingers, toes, and other projecting extremities. 4. Brown to black discoloration, skin having a leathery appearance.
Advanced decomposition	1. Moist decomposition with bone exposure less than one half that of the area being scored. 2. Mummification with bone exposure less than one half that of the area being scored.
Skeletonization	1. Bone exposure more than half the area scored, some decomposed tissue and body fluids remaining. 2. Bones largely dry, but retaining some grease. 3. Dry bone.

B.4 Gelderman et al.'s [17] decomposition scoring method for the head.

Table B.4: Gelderman et al.'s [17] scoring method for the head. Point.1, point.2, etc. represent the different phenomena.

Region	Score	Description
Head	1	1.1 No visible changes.
	2	2.1 Livor mortis, rigour mortis and vibices. 2.2 Eyes: cloudy and/or tache noir. 2.3 Discoloration: brownish shades particularly at the edges. Drying of nose, ears and lips.
	3	3.1 Grey to green discoloration. 3.2 Bloating of neck and face is present and/or skin blisters, skin slippage and/or marbling. 3.3 Purging of decompositional fluids out of ears, nose and mouth and/or brown to black discoloration.
	4	4.1 Caving in of the flesh and tissues of eyes and throat. Skin having a leathery appearance. 4.2 Partial skeletonization, joints still together.
	5	5.1 Gross skeletonization, some joints disarticulated.
	6	6.1 Complete skeletonization.

B.5 Gelderman et al.'s [17] decomposition scoring method for the torso.

Table B.5: Gelderman et al.'s [17] scoring method for the torso. Point.1, point.2, etc. represent the different phenomena.

Region	Score	Description
Torso	1	1.1 No visible changes.
	2	2.1 Livor mortis, rigour mortis and vibices.
	3	3.1 Grey to green discoloration.
		3.2 Bloating with green discoloration and/or skin blisters, skin slippage and/or marbling.
		3.3 Rectal purging of decompositional fluids.
		3.4 Post-bloating: release of abdominal gasses with discoloration changing from green to black.
	4	4.1 Decomposition of tissue producing sagging of flesh. Caving in of the abdominal cavity.
		4.2 Skin having a leathery appearance.
		4.3 Partial skeletonization, joints still together.
	5	5.1 Gross skeletonization, some joints disarticulated.
	6	6.1 Complete skeletonization.

B.6 Gelderman et al.'s [17] decomposition scoring method for the limbs.

Table B.6: Gelderman et al.'s [17] scoring method for the limbs. Point.1, point.2, etc. represent the different phenomena.

Region	Score	Description
Limbs	1	1.1 No visible changes.
	2	2.1 2.1 Livor mortis, rigour mortis.
		2.2 Discoloration: brownish shades particularly at the edges.
		Drying of fingers and toes.
	3	3.1 Skin blisters and/or skin slippage and/or marbling.
		3.2 Grey to green discoloration.
		3.3 Brown to black discoloration.
	4	4.1 Skin having a leathery appearance.
		4.2 Partial skeletonization, joints and tendons still together.
	5	5.1 Gross skeletonization, some joints disarticulated.
	6	6.1 Complete skeletonization.

C Chapter 5 Supporting Information

See Appendices B.4 to B.6 for Gelderman et al.'s [17] decomposition scoring method for the head, torso, and limbs, respectively.

Vita

Anna-Maria Nau was born in Austria. She moved to the United States in 2013 to compete as a Division I college athlete while earning a B.A. in Mathematics from California State University, Sacramento, in 2017. She then worked for 1.5 years as both a Healthcare Research Data Analyst at the UC Davis Medical Center, CA, and as a Data Scientist at Siemens in Vienna, Austria. In 2019, she began her Ph.D. in Data Science and Engineering at The University of Tennessee, Knoxville (UTK), working under Dr. Audris Mockus in the area of applied machine learning. During her doctoral studies at UTK, she completed a Data Science internship at Navy Federal Credit Union. In December 2024, Anna-Maria graduated with a Doctor of Philosophy degree in Data Science and Engineering. Her research interests include data processing, wrangling, and predictive modeling in the areas of machine learning, computer vision, and natural language processing across different industries.