PREDICTIONS OF INDENTATION STIFFNESS OF MUSCULOSKELETAL REGIONS USING ULTRASOUND

SEAN DOHERTY

Bachelor of Science in Biomedical Engineering

Duke University

May 2018

Submitted in partial fulfillment of requirements for the degree

MASTER OF SCIENCE IN MECHANICAL ENGINEERING

at the

CLEVELAND STATE UNIVERSITY

DECEMBER 2022

We hereby approve this thesis for

SEAN DOHERTY

Candidate for the Master of Science in Mechanical Engineering degree for the

Department of Mechanical Engineering

and the CLEVELAND STATE UNIVERSITY'S

College of Graduate Studies by

Committee Chairperson, Ahmet Erdemir

Thesis Chairperson

Department of Mechanical Engineering

Dr. Brian Davis

Thesis Committee Member

Department of Mechanical Engineering

Dr. Shawn Ryan

Thesis Committee Member

Department of Mathematics

Dr. Antonie van den Bogert

Thesis Committee Member

Department of Mechanical Engineering

Student's Date of Defense: December 6, 2022

ACKNOWLEDGEMENTS

Thank you to my family and friends for their unconditional support.

Thank you to Dr. Erdemir, Dr. van den Bogert, Dr. Davis, and Dr. Ryan for their valuable feedback and guidance on this project.

Thank you to my former coworkers at the Cleveland Clinic their prior work collecting the data that enabled my thesis.

Thank you to the funding sources, “Reference Models for Multi-Layer Tissue Structures” and the U.S. Army Medical Research & Materiel Command for their support of prior research fundamental to this work.

PREDICTIONS OF INDENTATION STIFFNESS OF MUSCULOSKELETAL REGIONS USING ULTRASOUND

SEAN DOHERTY

ABSTRACT

Tissue indentation response is an important metric for understanding how different musculoskeletal regions respond to loading and is a function of the tissue's form. Modern imaging techniques provide information about the internal structures of human tissue. Ultrasound remains one of the most common imaging techniques performed, given its portability and low costs. Prior work and data collection on 100 patients involved the collection of ultrasound images at eight different locations across the musculoskeletal extremities. Given the tissue structure information that the medical imaging provided, it was hypothesized that the mechanical properties of the tissue could be predicted from this data. This work aimed to incorporate various forms of patient data into different machine learning models for the prediction of tissue indentation response. These surrogate models would be capable of prediction of tissue compliance once input features are provided, potentially making them relevant in the clinical domain. Eight different surrogate models were developed, with four statistics models built and four deep learning models built to assess which method and which input factors were most suitable for accurately predicting indentation mechanics. The first four models were informed by tissue thicknesses and indentation region. The statistics surrogate models consist of two pure statistical models, while the other two models were based on a physics-based interpretation of two springs in series. The statistical models showed reasonable capability of predicting tissue surface stiffness, with the mean absolute percent difference

ranging from 25.4% to 29.7% across the four models. The deep learning approach was divided between two separate forms of deep learning. The first model was fed only demographic features, while a second model of demographics and manually extracted tissue thicknesses. These models also showed reasonable capability of predicting tissue indentation stiffness, with a mean absolute percent difference of 25.5% and 26.3%, respectively. A final modeling approach involved using convolutional neural networks, which utilized the raw ultrasound images. One model was only given the ultrasound image and gave a mean absolute percent difference of 31.5%. A final model consisted of the raw image, image metadata, and demographics and returned a mean absolute percent difference of 25.9%.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

1.1 Background

Imaging of patients provides one of the most important tools for healthcare, as it allows for insight on internal structures non-invasively. Imaging is often ordered by physicians to provide assistance in diagnosis and pre-surgical planning. The rate of imaging orders has increased rapidly since 1990 (Winder et al., 2021). This increase in imaging orders has tilted the cost versus benefit analysis of imaging, as increased imaging is largely performed on patients who do not require it, which leads to higher institutional costs. Given the high costs of certain scans such as magnetic resonance imaging for both a hospital and a patient, cheaper alternatives become more desirable. One of the cheapest imaging tools, ultrasound, has become the front line defense in patient imaging (Brattain et al., 2018).

Ultrasound is valuable for its high portability, low costs, and no need for patient shielding from harmful effects such as radiation (Liu et al., 2019). Additional benefits of ultrasound include it usage of conventional power sources and its ability to provide real-time images. These features are highly attractive in low income countries or highly rural areas where healthcare infrastructure is lacking or not present (Sippel et al., 2011).

Ultrasound has applications across the human body, such as within the musculoskeletal domain, viewing blood vessels, or observing a fetus (Brattain et al., 2018). While providing many benefits, ultrasound does have some key drawbacks. One of the main problems is that the signal to noise ratio of images tends to be quite low, especially as the ultrasound waves propagate through larger depths of biological tissue (Lento and Primack, 2007). This stems from the low variance in speed of sound between different tissue types. Ultrasound also cannot propagate through bone. A sample ultrasound image from the musculoskeletal domain, specifically the anterior upper leg, is shown, for visualization of the grainy nature associated with ultrasound (Figure 1) (Neumann et al., 2018). There is also the potential for high inter-operator variability or interpretation of an image, although this appears to be domain specific (Barbieri et al., 2008; Hadda et al., 2017; Lee et al., 2018). Other concerns involve the higher rates of wrist injury amongst sonographers, the people who perform the imaging (Coffin, 2014).

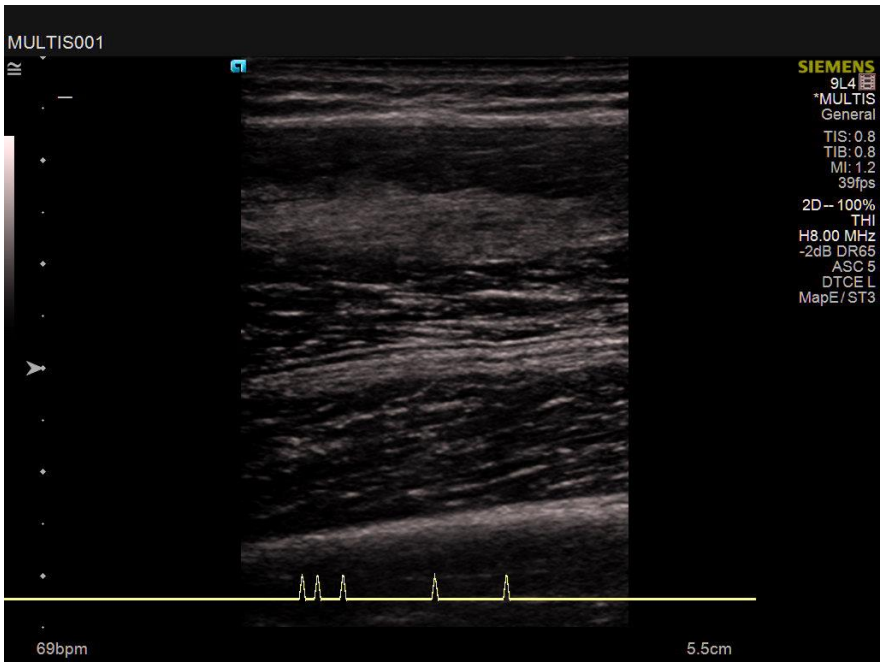

Figure 1: Ultrasound image of the anterior upper leg. *One of the main drawbacks of ultrasound is how grainy images appear. Spatial resolution is actually high, but images suffer from low contrast.*

While ultrasound provides image on tissue structures, how those structures mechanically behave or their health status is ultimately what is of interest. Extensible technologies, such as application of shear waves during ultrasound imaging, also known as ultrasound elastography, provide examples of how ultrasound can be used in detection of liver fibrosis or breast lesions (Sigrist et al., 2017). These studies provide localized tissue mechanics, but for overall force feedback the probe must be fitted with load transducers (Gilbertson and Anthony, 2013; Schimmoeller et al., 2018). Understanding how mechanical loads influence tissue response at the surface can provide valuable information on how internal tissue deforms. This information can provide value in surgical simulations (Satava, 1993), mitigating pressure ulcers (McInnes et al., 2015), prosthetic design (Faustini et al., 2006), or garment design (Wang et al., 2016).

While ultrasound has applications across human anatomy, this work is focused on its application in the musculoskeletal domain. The musculoskeletal extremities consist of both the posterior and anterior regions of the arms and legs. The leg region can be further subdivided into the upper leg (tissue close in proximity to the femur for both the anterior and posterior indentation) and the lower leg (tissue close in proximity to the tibia for anterior indentation and close to the fibula for posterior indentation). The same separation can be applied to the arm, where the upper arm can be defined by tissue near the humerus while the lower arm is the tissue surrounding the radius. The musculoskeletal extremities are a frequent location of injury, especially during military combat (Eskridge et al., 2012). These regions are exposed and not easily covered with body armor without restriction to mobility. These regions can be studied together given that their structure is highly similar (Figure 2). Each region is composed of a multi-layer tissue structure,

consisting of muscle attached to bone, which is covered by a protective layer of fat and an outward layer of skin. The intersection of ultrasound and the musculoskeletal system has proven applications in healthcare for the diagnosis of musculoskeletal conditions (Lento and Primack, 2007).

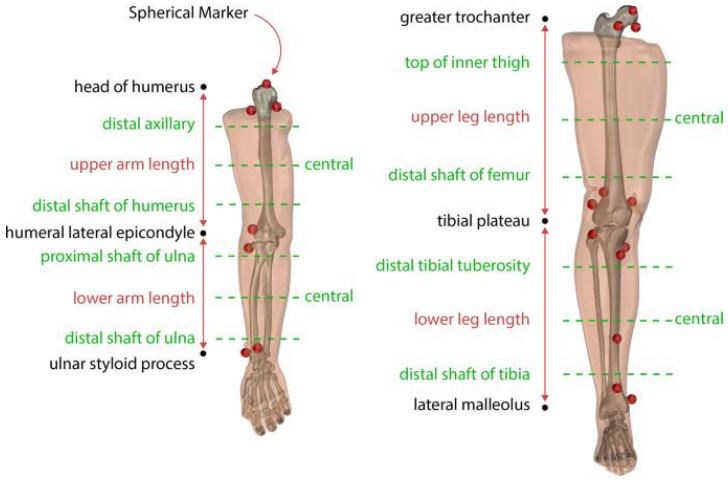

Figure 2: Anatomical depiction of musculoskeletal extremities. *A visualization of the regions of interest. Adapted from Schimmoeller et al. (2020) under Creative Commons license.*

The response of the multi-layer structure of the musculoskeletal extremities to surface loads can be defined as surface stiffness. The inverse of surface stiffness would then be surface compliance, a measure of how easy a material will stretch. Surface compliance could prove clinically relevant, it is important to understand what variables lead to differences in tissue response. Demographic information may be one source of data as an implicit indicator of the impact of biological variations that may affect surface stiffness. The significance of demographics, or lack thereof, is not agreed upon by the scientific community in several studies focused on tissue mechanics. For instance, two studies reported disagreement on whether age was a significant factor in indentation response of soft tissue (Choi et al., 2015; Neumann et al., 2019). Factors such as age intuitively may seem to impact soft tissue surface stiffness, as skin elasticity is known to

decrease as humans age due to the degradation of collagen over time (Farage et al., 2013). This decrease in elasticity could have an effect on surface compliance, although further comprehensive experimentation should be explored. For instance, a model generated solely based on age would omit key interplay between demographic information, such as the fact that due to lower estrogen levels, women's collagen begins to degrade at a younger age than for men (Farage et al., 2013). Similarly, ethnic background could play a role in skin biomechanical response as skin has been shown to have some different physical properties based upon ethnicity (Regueira et al., 2019; Wesley and Maibach, 2003).The above points focus on skin, but fat, muscle, and fascia could have similar differences across demographics. Changes in these tissues are more difficult to observe and measure than changes to skin. Change in skin structure can be seen with wrinkles, while changes in muscle can go largely unobserved. When attempting individualized models of soft tissue response, exclusion of demographic information may fail to capture the distinctness of each patient.

While demographics provide valuable information on a patient, the actual architecture and geometry of tissue and its distribution should also be considered in models. It is well established across engineering and biology that structure predicts function (Benjamin et al., 2008; Fithian et al., 1990; Knothe Tate et al., 2016; Li et al., 2019; Speck and Burgert, 2011; Tang et al., 2009). This raises questions about how the distribution of muscle, fat, and skin at different regions of the body affects indentation response. Li et al. highlighted this, by showing the surface mechanics of shale, a multi-component heterogeneous material, are dependent on the proportion and locations of the underlying constituent materials (Li et al., 2019). Individualized models should therefore

account for the structure of the underlying tissue, which ultrasound imaging can provide. Information on the thickness of each tissue layer at an indentation site can be observed from an ultrasound image.

Whereas Li et al. was able to assume literature properties for the underlying materials, doing so for biological models would likely lead to high levels of inaccuracy, where patient variability must be accounted for in models (Doherty et al., 2022; Moerman et al., 2017). Physics-based models, such as finite element analysis suffer from the requirement of accurate material properties. If the true response is known, inverse finite element can be calibrated to find accurate material parameters but requires multiple iterations of simulation for convergence. Finite element analysis also requires mesh sensitivity analysis, long compute times, and often human intervention during the construction of the model (Figure 3). Surrogate modeling aims to predict the same outputs, but in a data driven manner that does not require the same manual effort and runtime that a finite element model requires for a new prediction.

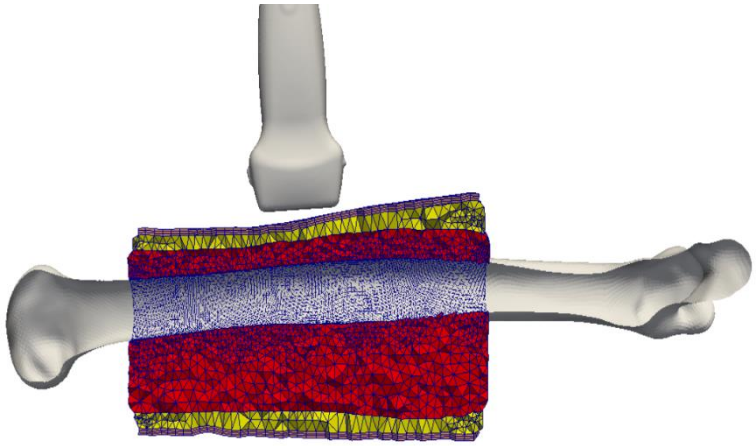

Figure 3: Sample finite element model. *Layered and 3D representation of the upper leg region with ultrasound probe. Meshing, material parameter optimization and mesh convergence make these simulations labor and time intensive.*

Two main forms of surrogate models are noted in the context of this study, both of which have proven capability in the biological and healthcare domains. Linear mixed effect models are a valuable tool in statistical modeling as they allow for accounting for effects from grouped populations amongst causal variables (Bernal-Rusiel et al., 2013; Gasparini et al., 2020; James et al., 2005). Linear mixed effect modeling is valuable because the significance of each variable can be determined, and the model's predictions are easy to understand.

The second form of surrogate modeling that will be used is different forms of deep learning. Deep learning applications in healthcare have rapidly grown across healthcare (Esteva et al., 2019). These models have high variance, meaning they are capable of learning many nonlinear trends that are located in a training dataset. A linear model may miss some of these key trends. A typical layout for one of the most common deep neural networks, a multi-layer perceptron, is provided in Figure 4. Datasets must typically be in the tens of thousands or even millions large for trends to accurately be identified. The advent of storing data digitally and accessible formats has created datasets that allow a machine to learn to become specialized at a specific task (Goodfellow et al., 2016). Deep learning is highly extensible though which makes it attractive for many types of problems or datasets. It has shown proven ability to analyze images, as well as other data including natural language and numeric data (Chollet, 2017; Shickel et al., 2018). Deep learning has surpassed human performance in some tasks, but adoption in healthcare has been slower than other industries. Healthcare data has many regulations surrounding it focused on important questions of data ownership and privacy, and data that is available is often less structured than traditional databases. Healthcare data

requires both domain expertise and data wrangling, as biomedical data is often poorly labeled, unstructured, and highly heterogeneous. Despite these challenges, artificial intelligence remains the best prospect for reforming healthcare and establishing a personalized medicine dream that many have touted for years (Topol, 2019).

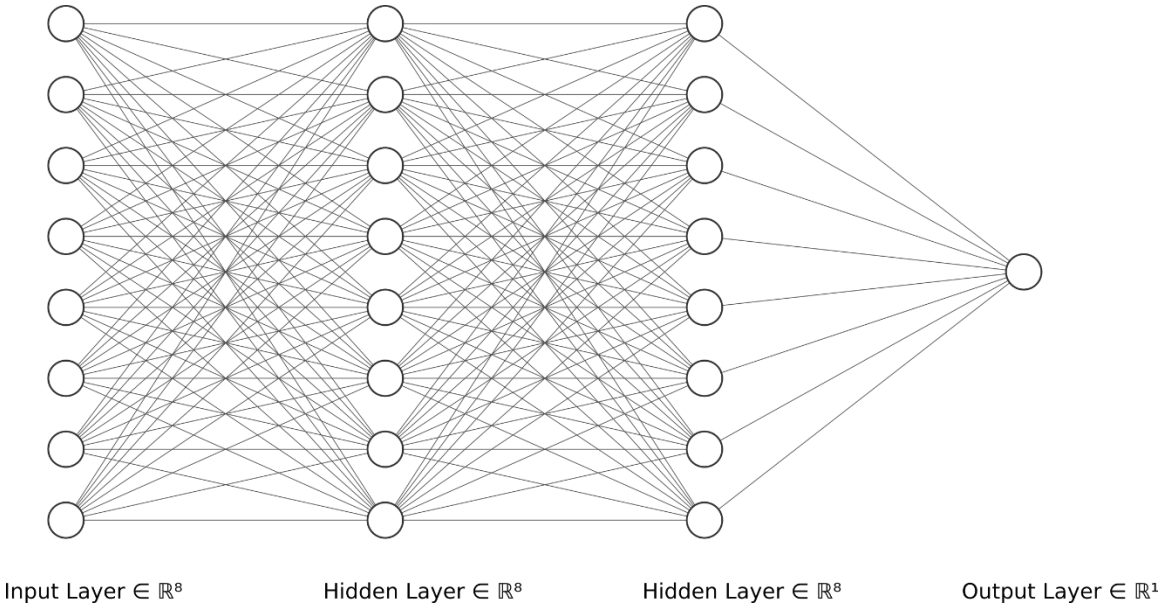

Figure 4: Sample neural network architecture. *This deep neural network has an input layer that takes 8 different input values, has two hidden layers, and an output layer for prediction of 1 value.*

1.2 Scope of Thesis

This thesis will aim to explore the usage of several different forms of surrogate modeling in an effort to predict soft tissue mechanics in the musculoskeletal extremities. Various data driven surrogate models will employ statistics and deep learning to make predictions on tissue compliance. Models will be fed different sources of input data based upon their formulation to see which models are suitable for this problem. Chapter I aimed to establish the motivation for exploring this area of biomechanics by providing a brief overview of soft tissue mechanics, clinical relevance, and surrogate modeling techniques.

Chapter II provides a detailed overview of the data that were used, how they were collected, and how they were processed. While these data were collected from a prior research project, it is important to establish the structure and available data to facilitate surrogate modeling.

Chapter III delves into how the first form of surrogate models were formulated. The mathematical basics and the underlying concepts that explain how the linear models' function. The chapter begins with the linear mixed effect models that were formulated as a part of prior work. The fundamentals of a linear mixed effect model and a linear model are laid out, before the mass-spring based formulation of some of the models is explained. Model results are then shown and discussed.

Chapter IV highlights the first portion of the project that incorporated machine learning. The chapter begins with an overview on how a neural network functions, particularly with respect to a deep neural network. Model structure and parameters are described and justified, for scientific reproducibility. An explanation of the methods to tune the network's hyperparameters are also be established. Model results and discussion are provided for two deep neural networks. Feature importance and deployment of the machine learning model concludes Chapter IV.

Chapter V continues the discussion of the application of machine learning to the realm of tissue mechanics through convolutional neural networks, this time relying on raw images. The underlying basics and how these networks differ from the previously used deep neural networks are established. Two different model architectures are shown, with subsequent results and discussion.

Chapter VI provides a brief summary of all the models created. It elaborates on the value of each model and why each model could or could not be chosen for usage. Conclusions from the work are laid out. The thesis wraps up with suggested future improvements or continuations that can stem from this thesis' methods and results.

CHAPTER II

BIOMECHANICAL DATA COLLECTION AND USAGE

Data lie at the foundation of any model built, but this statement is especially true for surrogate modeling. Surrogate models rely entirely on data to predict an outcome with no knowledge of the underlying rules that govern a system. Given this, it is critical to understand one's data when generating data driven models. An overview of data collection methods, data processing, and data distribution follows. Consideration of sources of bias that may alter a model's performance is also a key aspect of data driven modeling that are highlighted.

This thesis relies on data from "Reference Models for Multi-Layer Tissue Structures" (MuLTiS) a project that aimed to provide open-source *in vivo* and *in vitro* data for building of physics-based models (Erdemir, 2019). The *in vivo* data is of primary interest for this work. Data consists of ultrasound imaging and corresponding indentation testing for 100 different subjects at 8 different regions (Neumann et al., 2018). The regions consist of both the upper and lower arm and leg. Each region was tested in on

both the posterior and anterior sections of the tissue. Indentation trials were performed with an ultrasound probe that was fitted with a load transducer to accurately obtain forces associated with a specific image (Schimmoeller et al., 2019). The probe was not fitted for displacement tracking, rather displacement was measured by the change in tissue thickness. This force-displacement data is central to understanding tissue indentation response and stiffness. As Figure 5 shows, anatomical measurements were captured at other regions, but only the indentation sites (red dots) are of interest given their force-displacement data. Images were further processed manually by individuals, where the layers were demarcated and thickness measurements made based on these annotations.

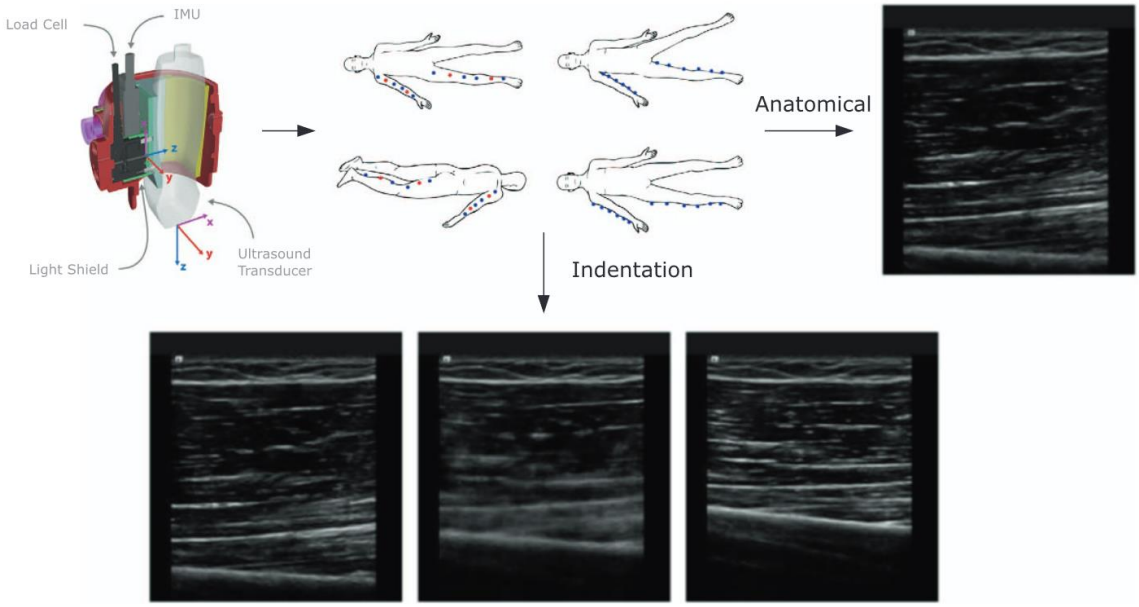

Figure 5: Workflow of experimental data collection. *Adapted from Neumann et al. 2018 under the Creative Commons license. Red dots show the points of indentation, while the blue dots show regions with only anatomical measurements.*

The demographic information is important for the models built, with a total of 100 research subjects tested. The participant pool consisted of 50 females and 50 males each tested at 8 sites, producing 800 total indentation trials. Error in experimentation led to the exclusion of 5 subjects, consisting of 3 women and 2 men. 8 individual indentation trials

also had to be excluded due to poor image quality that made manual annotation data too unreliable to trust. This left a total of 752 data entries summarized by gender and indentation location in Table 1. All locations had at least 45 samples for each gender. Location is defined by a 3-letter abbreviation, where the first letter denotes upper or lower (U vs L). The second letter denotes leg or arm (L vs A). An underscore then separates the indentation region as posterior vs anterior (P vs A). As a sample, UL_A indicates upper leg anterior while, LA_P represents lower arm posterior.

Table 1: Overview of 752 ultrasound images and indentation trials. *Location abbreviations represent region of indentation trails.*

Gender	**Location**	**Remaining Data Samples (50 originally)**
Female	LA_A	48
Female	LA_P	47
Female	LL_A	48
Female	LL_P	47
Female	UA_A	48
Female	UA_P	48
Female	UL_A	48
Female	UL_P	46
Male	LA_A	47
Male	LA_P	47
Male	LL_A	46
Male	LL_P	46
Male	UA_A	47
Male	UA_P	47
Male	UL_A	47
Male	UL_P	45

Data was collected from a wide background of individuals, meant to be a representative sample of healthy individuals from the United States. Demographic information such as age, race, body mass index (BMI), activity level, and ethnicity were recorded. Age ranged from 20 to 82 years old, with a mean age of 44.1. BMI is a function

of a patient's weight and height described by the following simple formula where kg is weight in kilograms and m is height in meters:

$$BMI = \frac{kg}{m^2} \quad \text{(Equation 1)}$$

BMI for the 752 data samples ranged from 17.27 to 44.98, with a mean BMI of 27.18. The shortcomings of BMI should also be noted, as BMI makes no distinction between muscle mass and fat mass. Race, ethnicity, and activity level were all self-reported by patients. No Asian or Black/African Americans identified as Hispanic or Latino. The number of data samples grouped by gender, race, and ethnicity is shown below.

Table 2: Distribution of race and ethnicity. *Note that the rows for Asian and Hispanic or Latino and Black/African and Hispanic or Latino are excluded since no subject identified as such. Data samples consists of images at any of the eight anatomical region and may include up to eight samples per patient.*

Gender	Race	Ethnicity	Number of Data Samples (All Regions)
Female	**Asian**	**Not Hispanic or Latino**	40
Female	**Black or African American**	**Not Hispanic or Latino**	94
Female	**White**	**Hispanic or Latino**	24
Female	**White**	**Not Hispanic or Latino**	222
Male	**Asian**	**Not Hispanic or Latino**	71
Male	**Black or African American**	**Not Hispanic or Latino**	62
Male	**White**	**Hispanic or Latino**	24
Male	**White**	**Not Hispanic or Latino**	215

Activity level was rated on a five point scale, self-reported by patients, although no patient reported the lowest or highest level. At the lowest level was extremely inactive, corresponding to someone unable to perform regular activities or someone completely bedridden. Sedentary was the next level, corresponding to a desk worker with little to no exercise (under 5000 steps a day). Moderately active was the middle level

(5000-10000 steps a day) followed by active (10000-12500 steps a day). As stated above, the highest category was extremely active, corresponding to competitive athletes or military personnel. The lack of participants in the extreme categories of activity may make BMI a reasonable metric, as BMI was originally formulated to describe average "healthy" individuals (Blackburn and Jacobs, 2014). As Table 3 shows, female respondents were mostly moderately active, while males were more likely to report sedentary or active behavior.

Table 3: Distribution of activity level. *No subject reported as being extremely active or extremely inactive. Data samples consists of images at any anatomical region and may include up to 8 samples per patient.*

Gender	Activity Level	Number of Data Samples
Female	**Active**	87
Female	**Moderately active**	255
Female	**Sedentary**	38
Male	**Active**	168
Male	**Moderately active**	141
Male	**Sedentary**	63

Analyzing datasets for bias is one of the ten key rules for proper analysis in machine learning (or other forms of data driven modeling) (Volovici et al., 2022). Several sources of bias exist and are readily apparent in the data. These potential biases could stem from the source of funding. As a Department of Defense project, there may have been more interest in capturing a sample that represents active military personnel rather than the population as a whole. The nature of performing human subjects research likely also plays a role. The ages of the subjects range from 20-82, notably excluding children and very elderly individuals. Any model predictions made on new subjects outside this range would be unreliable, a potential problem given this data is not representative of the population as a whole. United States 2020 census data shows that 22.2% of individuals

are under the age of 18 (“U.S. Census Bureau QuickFacts,” n.d.). Research testing on children is scrutinized very closely, making adding these subjects to the research study a justified omission. Statistics are not available on individuals over 82, but the over 65 demographic is 16.8% of the United States population. The dataset from MuLTiS contains 9 different individuals over 65 out of the 95 included subjects, or only 9.4% of the subjects, showing an under representation of elderly Americans.

Race and ethnicity data from the 95 subjects showed some deviation from the United States population as a whole, with Asians representing 14.7% of the study population and only 6.1% of the United States. Hispanic or Latino members were underrepresented by a factor of three in the study, with only 6/95 (6.3%) of subjects identifying as Hispanic or Latino when 18.9% of the United States identified as such during the 2020 census. Some of this discrepancy may stem from the fact the population in MuLTiS was meant to represent a population closer to military service members, but discrepancies such as these may limit the ability to draw trends from demographics for underrepresented populations.

With 95 subjects of usable data, the population groups become quite small as subdivisions are made on different minority groups. This makes data driven models likely to suffer from the ability to draw trends from the data, as the risk of overfitting increases greatly on small datasets. For example, no Hispanic or Latino individual identified as sedentary with respect to activity level, making the ability to draw conclusions about the interaction between ethnicity and activity levels tenuous on underrepresented groups. Similar biases exist across the data, such as the complete lack of minority groups above the age of 65. All 9 members of the over 65 group were white and not Asian or African

American, and none of the 9 members were also Hispanic or Latino. The oldest Asian person in the study was only 48. No Native American or Pacific Islanders were included in the study, nor were there any mixed races reported as well (perhaps due to lack of ability to specify multiple races on forms).

Modeling, simulation, and machine learning have developed rapidly in industry and businesses, where data tends to be cleaner, larger, and less regulated. Healthcare has lagged in adopting computational models and machine learning given the challenges of working with biomedical data (Erdemir et al., 2020). Data issues such as the sources of bias discussed in this Chapter make it important to discuss the context of use for the surrogate models that follow in this thesis. Context of use defines what decisions a model will drive with its results. A model that drives a surgeon's decision or national public health policy needs a high level of scrutiny and validation. The models for this project fall on the lower left corner of Figure 6 because they have low use capacity, meaning they have a limited range of usage. These models exist in the hypothesis testing realm and could not be used for any sort of translational or clinical research in their current state due to the data issues described above, and the small amount of data. Expanding the amount of data would be imperative before the model predictions could be trusted with a decision of any clinical importance in a real-world hospital setting. The following chapters will discuss the decisions made for model selection, but model quality is considered much less important than data quality. A popular saying in machine learning fields is that "A dumb algorithm with lots and lots of data beats a clever one with modest amounts of it" (Domingos, 2012). Readers of this thesis should keep this idea at the forefront of their mind when reading the remaining chapters. For individuals interested in

using the following models, it needs to be clear that the models exist in the realm of hypothesis data exploration and fundamental research. The models should not be used outside of this domain of use unless significantly more data are collected.

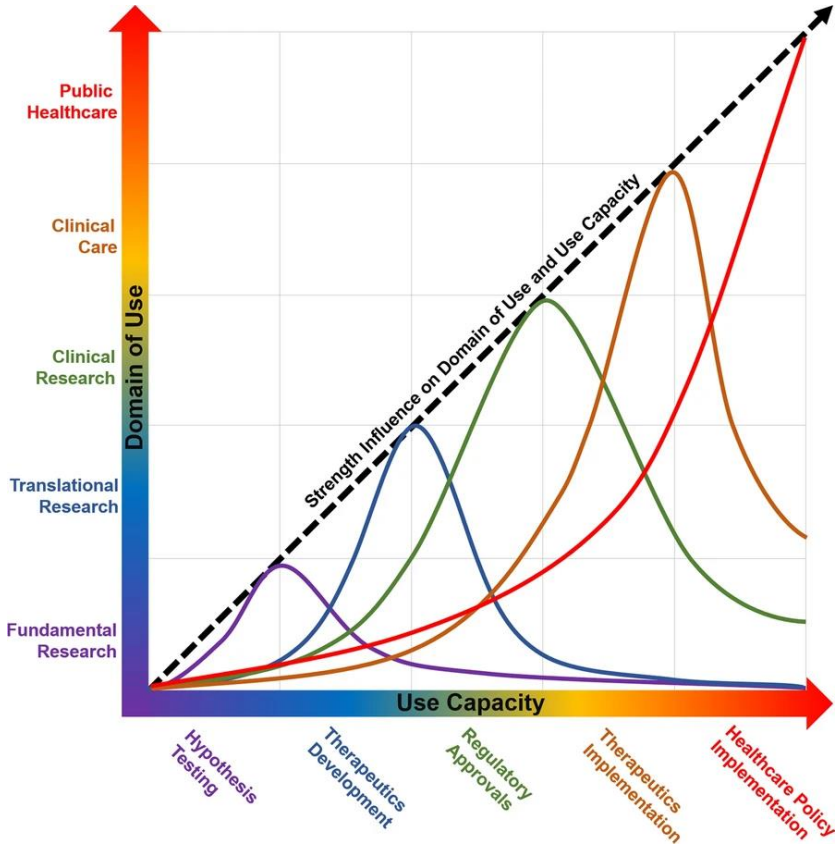

Figure 6: Context of use curves. *Adapted from Erdemir et al. 2020 under the Creative Commons License. The surrogate models developed would fall on the lower left corner.*

With the demographic information established, establishment of additional model input features is the next important data topic. Three additional features were generated, although only two made it into this thesis's models: fat, and muscle thickness. Skin thickness was omitted due to an unacceptable signal to noise ratio in thickness measurements. These markers were generated through manual annotation of ultrasound images by marking the location of different layers. Distance from each marker could then be measured as an estimate of each layer's thickness (Figure 7). These annotations and measurements were performed multiple times, by the same individual to analyze the

intra-observer variability, and by a different individual to measure inter-observer variability (Neumann et al., 2018). Both annotations were done blind (without reference to previous annotations) and 1 week apart for inter-observer measurements. Both individuals making the measurements were trained by an ultrasound radiologist. Total thickness measurements for both inter- and intra-observer variability had a mean absolute difference of 0.19 mm and 0.32 mm respectively. Variability was higher for individual layers, but still in a reasonable tolerance relative to overall thickness of each layer except for skin thickness. Mean absolute difference represented ~40% of skin thickness for inter-observer measurements, suggesting error may exceed actual variability in skin thickness and highlighting the need for skilled radiologists to provide these feature annotations if skin thickness is desired for models.

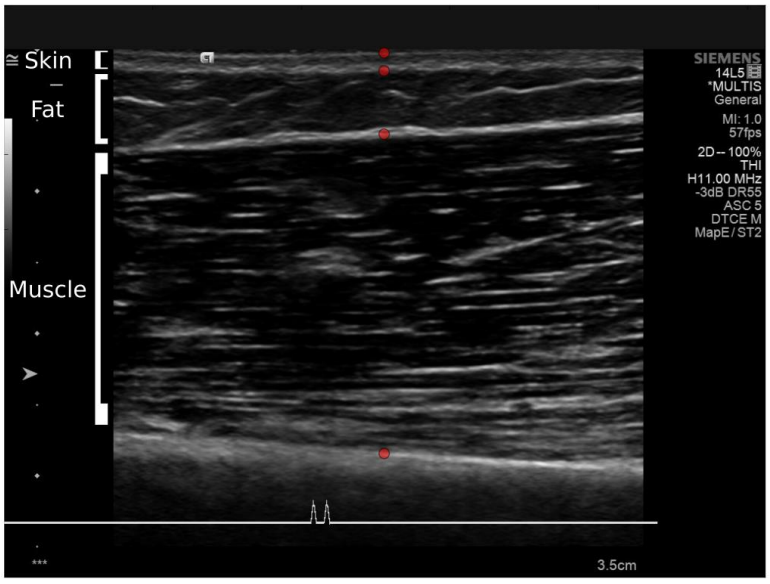

Figure 7: Manual annotation of tissue layers. *This ultrasound image comes from the anterior upper arm. The measurement of skin is highly sensitive to accurate positioning due to its small thickness.*

The final point of data that is important to discuss is the response variable for all of the surrogate models, tissue compliance. Tissue compliance is used over tissue stiffness as a matter of convenience, as it could be linearly related to tissue thickness (whereas tissue stiffness is inversely related). This is a metric to estimate how the bulk tissue responds to surface loads, as higher surface compliance means more deformation with similar loading. The direct measurement of surface stiffness can be defined as MPa/mm, which can be found with a linear fit of pressure (defined by force over an assumed constant probe area during indentation) over displacement (tissue thickness change from exterior skin to bone surface) (Neumann et al., 2019). Surface compliance can therefore be defined as the inverse of surface stiffness (mm/MPa). Normalizing displacement values by a region's total indentation thickness may be valuable for future work, but surface compliance is defined as above for consistency with prior work. Surface compliance will be predicted by the models in Chapters III, IV, and V. The relationship between tissue compliance and numeric features that will be used in modeling are shown in a scatter plot that compares each variable to each other (Figure 4). The diagonal portion of the plot is known as a kernel distribution estimation plot, which shows the frequency of that variable (i.e. the 1st row, 1st column plot shows the distribution frequency of age, highlighting the research population skews younger). Figure 4 highlights some of the challenges associated with working with biological patient data, where variability is high and trends are unclear.

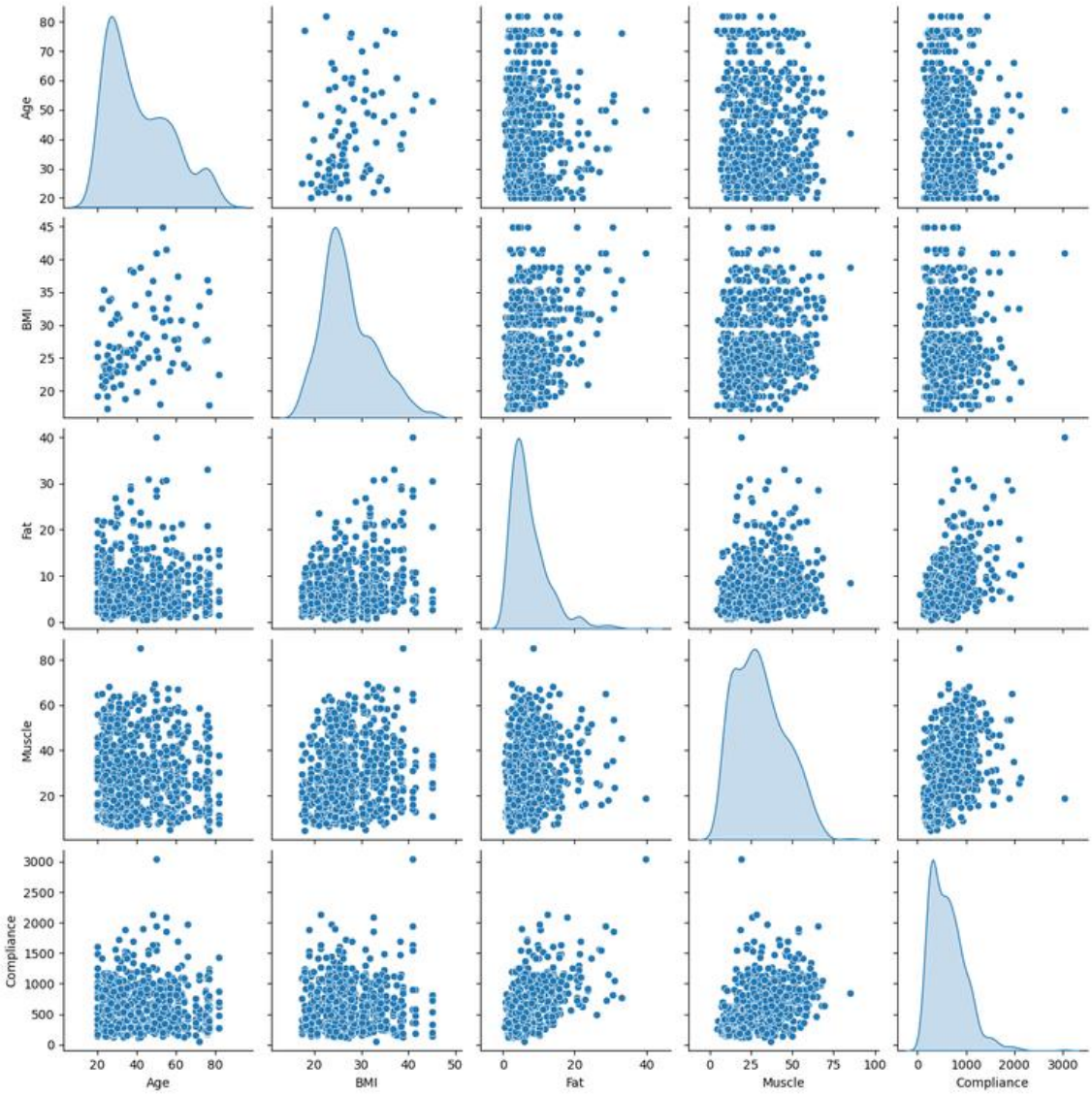

Figure 8: Pair plot of the numeric data in the models. *The diagonal column contains frequency distribution information, while the non-diagonal shows a scatter plot relating the row variable to the column variable.*

A reasonable question for measurement of surface compliance is that soft tissue deformation response is considered nonlinear. Response is usually defined with a hyperelastic constitutive model (Mihai et al., 2015). Other factors such as anisotropy or viscoelasticity are also important considerations for more accurate modeling of tissue response. Performing a linear fit on the displacement data is therefore an approximation

made to simplify the representation of indentation response. This approximation also does a reasonable job of fitting the force-displacement data obtained, based on R^2 fit of the data. All 752 trials were fit with both a basic linear regression and non-linear regression of the form:

$y = ax^2 + bx$ (Equation 2)

Of the 752 data samples, 735 samples, or 97.7% of the data, had an R^2 value difference of 0.1 or less between the non-linear regression and the linear regression (Neumann et al., 2019). This shows the approximation is sufficient for a fundamental analysis exploration (Figure 9). A model for clinical use may consider predicting a hyperelastic constitutive model's coefficients. Given a linear fit, it may be possible to reverse engineer a more accurate nonlinear fit as well, as in Figure 9 the characteristic behavior of the linear model under predicting pressure at higher displacements can be observed. It is worth considering how repeatable these displacement and pressure measurements are, which previous experimental data did not assess. The variability in experimental data may influence the lower bound of how accurate a model can be.

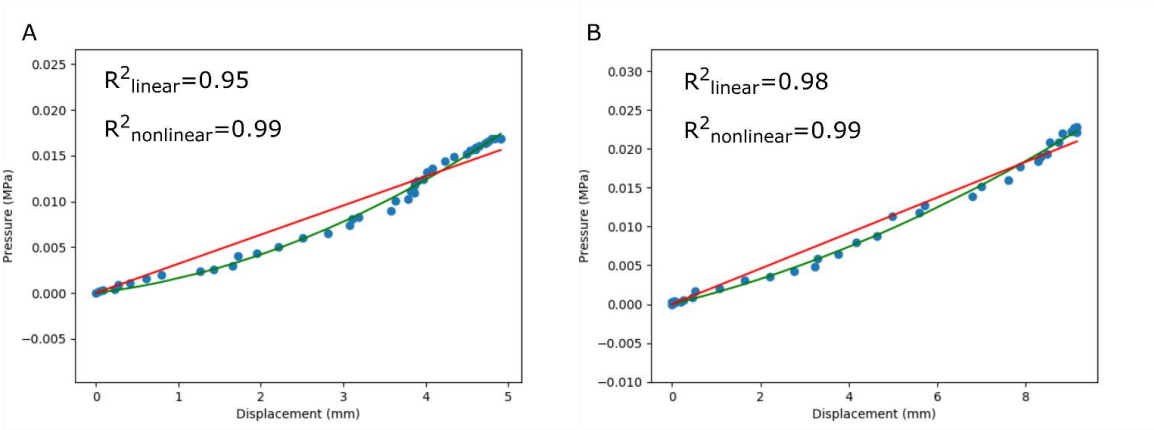

Figure 9: Sample of reasonable linear fits. *Adapted from Neumann et al., 2019 under the Creative Commons license. The nonlinear fit more accurately fits the data, but the linear regression is easier to predict and sufficient for this thesis.*

CHAPTER III

LINEAR MODELING OF TISSUE COMPLIANCE

The power of machine learning has led to its proliferation across domains. Calls for restraint in its use have also been made, as Volovici et al. proposed more careful usage of machine learning for clinical research (Volovici et al., 2022). In their work, 10 rules were proposed to properly use artificial intelligence with clinical data, with the first rule being that traditional statistical methods should be used alongside machine learning based methods. These traditional statistical methods may end up being as accurate as the machine learning methods on certain data and are less of a black box model than machine learning is.

When working with clinical data, it is often the case that the data contains different populations of individuals. These different populations may receive a mixture of treatments, requiring a more sophisticated statistical method than a simple linear regression to handle this categorical data. Linear mixed effect models and multiple basic linear models are one such approach for this sort of data (Faraway, 2005). These models will predict surface compliance (mm/MPa). The analysis in this chapter is based on prior analysis by the author and several coworkers, in a forthcoming publication (Neumann et al., in review). The prior work was performed and modified for this thesis using R version 3.4.2 (R Core Team, 2017).

Two different forms of variables or effects can be used in a linear mixed effect model. The first is a fixed effect, which is similar to a dependent variable. It is expected that a fixed effect will have an influence on surface compliance. Only two of the four models developed were linear mixed effect models, while the other two models were linear models. For the linear mixed effect models, there were two fixed effects and two random effects. Both fixed effects came from the annotated ultrasound images, initial fat thickness and initial muscle thickness. Fat and muscle thickness are used, because it was hypothesized that tissue response is a function of layer thickness. It is important to note again that skin thickness is not used in the model because of the lack of repeatability in measuring skin thickness. Any modeling of skin thickness may just be fitting to noise. The random effect in the two linear mixed effect models is a form of categorical data that is used to group data. In the two linear mixed effect models, location and subject ID (a number to distinguish different patients) is used as a random effect. This was a single model and can be labeled as the combined location model. In the other two models, which are just linear models, there are no grouping categories and the model only sees the initial tissue thicknesses. The models were however formulated specific to region; eight different models only saw the data from one of the eight regions. In this manner, the influence of patient variability could be examined. Model performance was measured with mean absolute percent error, a measure of how much a predicted value deviates from the experimental value, divided by the experimental value.

The models were also split into two types, with each type of model receiving one of the two random effect treatments listed above (location specific and combined location. The first model has an intercept and is labeled as the statistics based model. The

coefficients here have no physical intuition, as having a non-zero pressure associated with zero displacement is illogical. To provide physical meaning, a second model was developed where the model's y-intercept was forced through zero. The coefficients of this model are then analogous to spring coefficients of both fat and muscle, making the model representative of two springs in series (Figure 10). The slope of this line represents the inverse of Young's Modulus.

Table 4: Model table for statistical models.

Model Name	Intercept at (0, 0)?	Fixed Effects	Random Effects (for mixed effect models)	Number of linear regressions
Physics-based combined location	Yes	Initial fat and muscle thickness	Subject ID and indentation region	1 (all locations together)
Physics-based location specific	Yes	Initial fat and muscle thickness	N/A (linear model)	8 (separate for each location)
Statistical combined location	No	Initial fat and muscle thickness	Subject ID and indentation region	1 (all locations together)
Statistical location specific	No	Initial fat and muscle thickness	N/A (linear model)	8 (separate for each location)

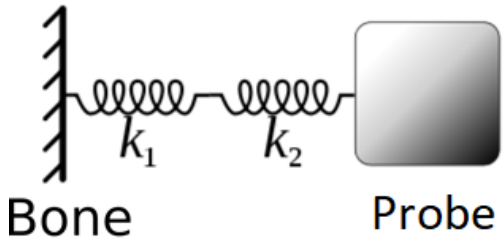

Figure 10: Physics-based model spring-mass system. *Adapted from Wikipedia under the Creative Commons license. k_1 is the spring stiffness for the muscle layer, while k_2 is the fat layer's spring stiffness. The rigid interface is the bone, while a force is applied to the mass of the ultrasound probe.*

The statistics-based models were provided with the 752 data points. Similar to the deep learning models in the next chapters, the data was subdivided into training and test data. 601 samples were in the training set (80% of the data), while 151 samples were in the test set (remaining 20%). Unlike the later chapters, there was no need for a validation data set, as a linear fixed effect model has no parameters to tune, unlike a deep learning model. As such, the validation data was not split off, and was instead included in training.

The first model was the statistics model with locations combined. The model had random effects of subject ID and location, and fixed effects of muscle thickness and fat thickness, extracted from manual annotation (Figure 7). Note that this figure does not show model predictions against input variables, but rather shows model outputs against experimental surface compliance values to assess goodness of fit. The model yielded a mean absolute percent error of 26.64% on the test data set (Figure 11). Coefficients can be found in Table 5.

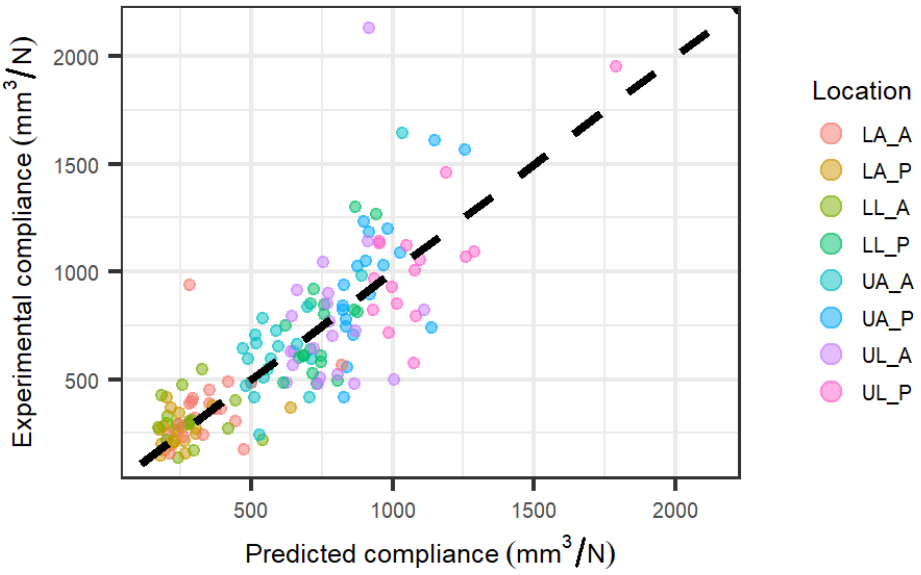

Figure 11: Statistics model (all locations), predicted vs experimental compliance. *Data points are colored by indentation region.*

Table 5: Statistical models coefficients. ** in the table refers to p-value < 0.05, ** refers to p-value < 0.01, *** refers to p-value < 0.001. Standard errors are listed in parentheses. Region abbreviations are defined in Chapter II.*

	LA_A	LA_P	LL_A	LL_P	UA_A	UA_P	UL_A	UL_P	Combined
Intercept	317.2 *** (44.1)	345.4 *** (55.7)	327.4 *** (55.6)	784.7 *** (175.8)	418.8 *** (100.5)	447.5 * (181.9)	354.7 ** (133.5)	1157.2 *** (220.1)	402.28 *** (92.0)
Muscle	-2.49 (2.80)	-6.47 * (2.91)	0.12 (1.34)	-3.23 (3.28)	4.52 (3.48)	6.85 (4.94)	8.05 (4.13)	-3.55 (3.67)	2.93 * (1.27)
Fat	6.14 (4.10)	9.40 (9.09)	-3.74 (4.03)	12.25 (6.33)	9.31 (5.73)	28.11 *** (6.70)	15.24 ** (5.15)	8.35 (4.95)	18.05 *** (2.19)

The second model was the linear model with locations separated, which had only fixed effects of muscle and fat thickness. This model produced a mean absolute percent error of 25.68%. The distribution of predictions in this model (Figure 12) is noticeably different than in Figure 11 for some regions. The anterior lower leg predictions are quite close to each other, as seen in the near vertical line of predicted vs experimental

compliance for this region. This suggests the anterior lower leg model has little discriminative value and is essentially only predicting the mean for the region. This is a valid strategy for reducing error, but shows the model is no better than literature values. At least for this region, fat and muscle thickness have nearly no predictive power in this region. The combination of the eight linear models exceeds the performance of the first statistical model, although only very slightly. This could stem from the fact linear mixed effect models have random intercepts for each factor, while the linear model reduces error by minimizing the intercept as well. The linear models do have no consideration of patient variability, so the introduction of subject ID is perhaps not significant enough in these indentation trials. A comparison of coefficients can be found in Table 5. The coefficients in Table 5 differ from the coefficients of prior work because they represent only the test dataset rather than the entire dataset (Neumann et al, in review).

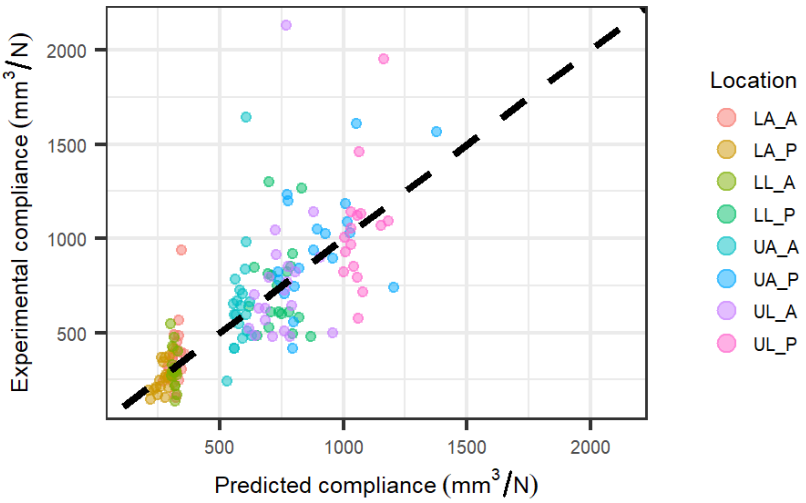

Figure 12: Linear model (location-specific), predicted vs experimental compliance. *Data points are colored by indentation region.*

The remaining two statistical models consist of the physics-based models. The model formulation was essentially identical to the above, with one linear mixed effect model and one linear model. The main difference for these models is the model is a

regression through 0, where the y-intercept is forced through (0, 0). Doing this allowed for a formulation where the coefficient of the model became equivalent to spring stiffness coefficients. Substituting Hooke's Law into the equation for Young's Modulus, the following formulation can be derived, where A is probe area (mm^2), k is overall stiffness (mm/MPa), t_m is muscle thickness and t_f is fat thickness in mm, while E_m and E_f are Young's modulus in MPa:

$$\frac{A}{k} = \frac{t_m}{E_m} + \frac{t_f}{E_f} \quad \text{(Equation 3)}$$

These coefficients could then be used to predict tissue displacement based on maximum probe forces, although this work is outside the scope of this thesis. For the physics-based combined location, the model had a mean absolute error of 26.73% (Figure 13). There is a minimal difference in between the predictions made by the physics-based model (Figure 13) and the statistics linear mixed effect model (Figure 11). This can lead to viewing the statistics model as a sensitivity test for the physics-based models, where forcing the model through zero had only a small impact on predictions.

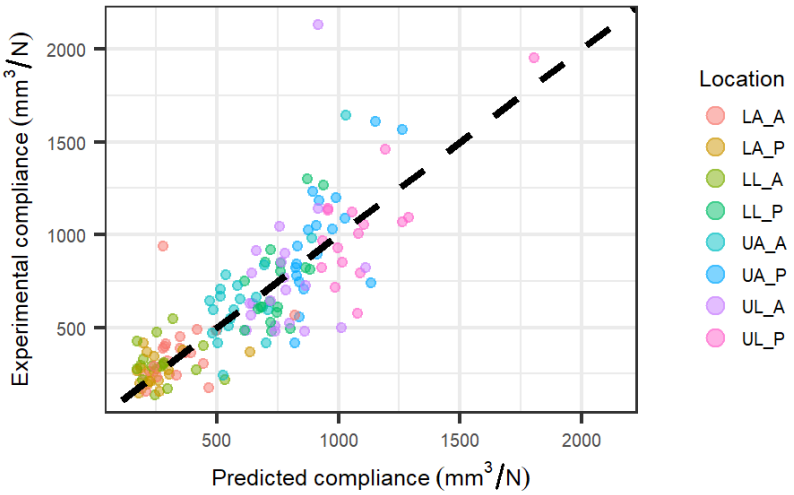

Figure 13: Physics-based model (all locations), predicted vs experimental compliance. *Data points are colored by indentation region.*

The final statistical model was the physics-based linear model (only fixed effect of initial muscle and fat thicknesses, random effects of location and subject ID). This was the worst performing of the four models, with a mean absolute error of 29.48%, perhaps because the linear model could only use minimize error through the slope (Figure 14). It should be noted though that the standard deviation on the mean absolute errors was nearly equal to the mean absolute error for each model, suggesting model performance ordering may change if the testing data consisted of a different set of data points. The prediction quality from each model is therefore more similar than they are different. The coefficients in Table 6 are noticeably different from Table 5, unsurprisingly since there is no y-intercept. Several of the coefficients in Table 5 are negative, indicating increasing initial thickness leads to decreasing compliance, while all coefficients in Table 6 are positive.

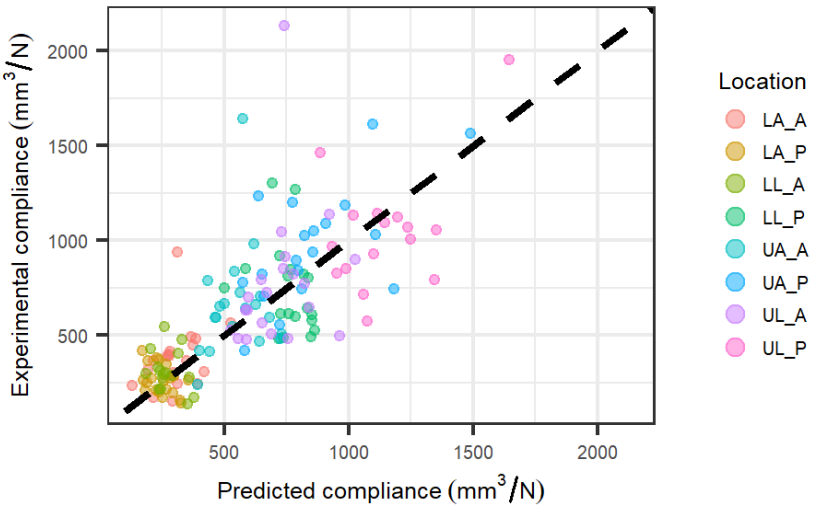

Figure 14: ***Physics-based linear model (location-specific), predicted vs experimental compliance.*** *Data points are colored by indentation region.*

Table 6: Physics-based model coefficients. *These models had no intercepts. * in the table refers to p-value < 0.05, ** refers to p-value < 0.01, *** refers to p-value < 0.001. Standard errors are listed in parentheses. Region abbreviations are defined in Chapter II.*

	LA_A	LA_P	LL_A	LL_P	UA_A	UA_P	UL_A	UL_P	Combined
Muscle	13.88 *** (2.12)	8.78 *** (1.91)	7.32 *** (0.66)	10.65 *** (1.15)	18.11 *** (17.97)	18.76 *** (2.24)	17.76 *** (1.99)	15.11 *** (1.10)	3.49 ** (1.25)
Fat	23.86 *** (4.26)	43.83 *** (8.87)	8.28 (4.18)	27.99 *** (5.90)	17.97 ** (5.91)	38.86 *** (5.25)	19.63 *** (5.08)	23.01 *** (4.77)	18.50 *** (2.17)

The above statistical models could be valuable tools, given several benefits of using linear models. The first valuable aspect of the models is the presence of coefficients with standard errors and p-values, meaning they are fully explainable. Model results would not change when training on the same data (this can also be true of deep learning, but requires fixing the value of a random number generator seed). Physically meaningful coefficients are also valuable, as they can be applied to additional modeling scenarios as they were in a forthcoming publication (Neumann et al, in review). Lastly, there is minimal overfitting risk with a linear model. These models do have drawbacks, namely that they consider only a few input features. Initial muscle and fat thickness also requires manual annotation, a process that takes time and some experience in interpreting ultrasound images. There are often streaks of brightness in ultrasound images based on probe angle that could be misinterpreted as a tissue delineation. The value of these models can ultimately be more accurately assessed through comparison against the models of Chapter IV and V.

CHAPTER IV

DEEP NEURAL NETWORK BASED PREDICTION OF INDENTATION RESPONSE FROM CATEGORICAL AND CONTINUOUS INPUTS

Deep neural networks, also known as deep feedforward networks or multi-layer perceptrons, have achieved rapid adoption across industries due their innate ability to work with higher dimensional data that more simple statistical learning methods are incapable of processing at once. These networks represent the simplest and earliest developed neural network architectures (Schmidhuber, 2015). This model has exploded in popularity and achieved state of the art results across domains (Schmidhuber, 2015). These models are mainly used for performing non-convex optimization, where prediction of a global minimum is a challenging task. The models are known as feedforward, as information flows from the input to the output prediction with intermediate computations that has no feedback of the output back into the model. Information moving through the network is known as forward propagation, which produces a cost value that is calculated from the cost function. Back propagation can then occur, where information from the cost function flows from the output to the input, which provides a gradient for updating model weights. Model weights and biases are randomly set, so back propagation is a key step

that allows a model to go from random guesses to ones that accomplishes the sophistication to fit the training data.

The initial dataset as described had many intersecting trends that could not necessarily be analyzed together with a traditional statistical method like a mixed effect model. The forthcoming publication based on the linear modeling of surface compliance posited that machine learning would be able to unlock some of the deeper trends and associations in this dataset (Neumann et al., in review). Following the recommendations of Volovici et al., this chapter will lay out data processing decisions, architecture choices, hyperparameter searching, and model results (Volovici et al., 2022). Two separate deep neural networks were built in the Python library Tensorflow (Abadi et al., 2016). The first model was built entirely on demographics information and other categorical data, with an output of tissue surface compliance. The inputs to the model consist of all features located in the Chapter II data discussion (also summarized in Table 7).

Table 7: Inputs to the demographics-based deep neural network. *Indentation location is defined as a three letter abbreviation (first letter is upper or lower, the second letter as leg or arm, third letter after an underscore defines posterior vs anterior).*

Variable	Possible Data Values
Indentation Location	[UL_A, UL_P, LL_AC, LL_PC, UA_A, UA_P, LA_AC, LA_PC]
Gender	[Male, Female]
Activity Level	[Sedentary, Moderately Active, Active]
Race	[White, African American, Asian]
Ethnicity	[Not Hispanic or Latino, Hispanic or Latino]

BMI	17.27-44.98 (Float)
Age	20-82 (Integer)

The data is first processed by one-hot encoding, a method of removing ordinality from a list. In the above list of variables for example, rather than classifying a person in a column of race as white, Asian, or African American, new columns are made for each variable. Categorization is then defined by a binary 1 (true) or 0 (false). Someone who is Asian would have a 1 in the new column defined as Asian, while the White and African American new columns added would have a 0. The race column is then no longer required, given each of its responses has become its own column. Sample one-hot encoded subject data is included in Table 8.

Table 8: Sample subject data after one-hot encoding. *A 1 indicates this categorization is true, while a 0 indicates a categorization is false. Numeric data, such as age or BMI, are not one-hot encoded but are later normalized into the 0-1 range.*

Active	0
Moderately active	1
Sedentary	0
Hispanic or Latino	0
Not Hispanic or Latino	1
Female	0
Male	1
LA_A	0
LA_P	0
LL_A	0

LL_P	0
UA_A	0
UA_P	1
UL_A	0
UL_P	0
Asian	0
Black or African American	0
White	1
Age	31
BMI	31.13

A key aspect of machine learning is the separation of training, testing, and validation data. Given machine learning is a high variance form of modeling, it is easy for the model to over fit on the training set. While learning the training data very well seems desirable, this reduces the model's ability to generalize with new data it sees, and instead the model has purely memorized the training data inputs and outputs. This is analogous to fitting data points with a very high order interpolating polynomial, where the function will intersect all data points, but any new data would be poorly predicted due to the extreme oscillations in the function. To alleviate this, a random sample of 20% of the data was split into a test set, which is not shown to the model until training and model tuning are finalized. This can be considered a method to approximate the real-world performance of the model. 16% of the total data was also split into a validation data set, which is used for optimizing the model architecture (16% comes from 20% of the

remaining 80% of the data). This data set is used for optimization of model parameters, known as hyperparameter tuning, as well as indicating when training has begun to over fit. This data is not used for training in the model once hyperparameters are established, as this provides extra weight to this slice of data, where it is used for training while also being used for selection of optimal model parameters.

Another key step in manipulating data is normalizing the data. Data in its raw form as an input varies in size considerably, which can lead to gradients in the neural network diverging and producing poor convergence. This is addressed by scaling the data from 0-1. This normalization layer serves as the input layer to the neural network, with a number of inputs equal to the one-hot encoded vector length. Different hyperparameters alter model structure slightly, but the model generally is highly similar to Figure 4, where the input layer is followed by 2 hidden layers, followed by one output layer (a single neuron to predict a single floating point value for tissue compliance). Further structure of the model is defined through a hyperparameter search, with specific values located in Table 9. Hyperparameter space was explored over 20 iterations, with each model run twice and validation loss averaged between the two trials. The best hyperparameters were used in evaluating the test dataset. The hyperparameter search included the model's learning rate, which represents how large of a step size will be taken. It is important to vary the learning rate value, as too small of a learning rate may lead to getting stuck at a suboptimal local minima, while too large of a value can lead to overshooting the global minimum are poorer convergence. Another hyperparameter was the number of neurons in the hidden network. Too many neurons can lead to overfitting, while too few neurons will lead to underfitting the solution. The remaining three hyperparameters all represent

different techniques used to avoid overfitting the machine learning model. Gaussian noise introduces random augmentation to data values, to prevent the model from seeing the exact same inputs. Dropout rate functions by dropping a proportion of neurons between layers. This somewhat counterintuitive process can improve model results and reduce overfitting by ensuring that all nodes in the model are valuable, rather than only a select few. The final hyperparameter designed to reduce overfitting was L2 regularization. L2 regularization adds the squared value of the model weights into the loss function, as a method of keeping weights small to prevent large curvature changes in the approximated function.

Table 9: Deep neural network hyperparameter search. *Searching of values is conducted with a random search of parameters.*

Hyperparameter	Potential Values	Search method	Optimal value
Learning Rate	[1e-2, 1e-3, 1e-4]	Choice of value	1e-2
Number of Neurons per Hidden Layer	min = 32, max = 256, step size = 32	Sampling at step size intervals	128
Gaussian Noise	0.1	Boolean	True
L2 Regularization	min = 1e-5, max = 1e-2	Log sampling of solution space	2.23e-4
Dropout Rate	min = 0.1, max = 0.6, step size = 0.1	Sampling at step size intervals	0.5

It should be noted that Table 9 is not an exhaustive list of the inputs to the model. The model had a batch size, or how many data points the model should see before it updates its weights, of 16. Another key aspect of a neural network is an activation function, to introduce nonlinearity into the model, as without these functions the model can only fit linear behavior. The model in this work utilized a Parametric Rectified Linear Unit (PReLU) function between the neural network hidden layers. This activation function operates similar to a rectified linear unit (RELU), defined by:

$f(y_i) = y_i \qquad y_i > 0$ (Equation 4)

$f(y_i) = 0 \qquad y_i \leq 0$

The PReLU activation function aims to avoid one of the major issues of the RELU activation function, where any negative value will turn a neuron off and leads to an inability to update weights with backpropagation when the function returns 0 (He et al., 2015). The PReLU activation function is defined as:

$f(y_i) = y_i \qquad y_i > 0$ (Equation 5)

$f(y_i) = \lambda y_i \qquad y_i \leq 0$

The lambda value reduces the slope of the line for negative values as compared to positive values. This allows backpropagation to occur on the entire network. A value of 0.1 was used for this lambda in this model.

The model was compiled with mean absolute error as its loss function, with the Adam optimizer, a popular first order gradient-based optimization function (Kingma et al., 2020). The learning rate was reduced on when the validation loss plateaued for 20 consecutive epochs by a factor of 10. If no improvement was achieved for 50 epochs, the model training was stopped. 1000 epochs were prescribed for training, with each epoch

representing the model seeing the entire training dataset once. In practice, training usually halted before 100 epochs due to lack of improvement in validation loss, signaling overfitting was occurring. Optimal model weights were considered to be the weights that produced the lowest validation loss (mean absolute error). Model training was performed on an Nvidia RTX 2080 Super GPU, and took around 40 minutes to tune hyperparameters and then perform the final training and predictions. Model results were then deployed to a Gradio interface for possible user interaction (Abid et al., 2019). The Python library Shapley was used for insight into feature importance (Lundberg and Lee, 2017).

When evaluated on the test dataset, the demographics-only model reported a mean absolute error (MAE) of 158.42, and a mean absolute percent error (MAPE) of 25.53%. The training history of the model is shown in Figure 15. This curve shows the characteristic behavior of overfitting, where the network's loss continues to decline on the training set (blue line), while the validation loss actually starts to increase as epochs increase (orange line). This overfitting is not a concern, given that the model can simply load parameters from its earlier iterations before it started overfitting. The optimal validation loss was achieved at epoch 14, and training was terminated at epoch 64 (due to lack of reduction in validation loss on 50 consecutive epochs, the early stopping criteria), showing that reduction of the learning rate after 20 consecutive epochs of no validation loss reduction did not help the model achieve better performance.

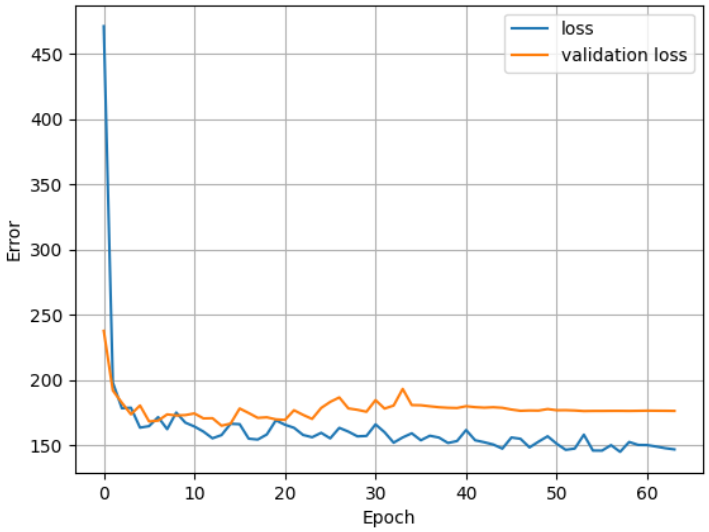

Figure 15: Demographic model MAE training history. *The validation loss was minimized at epoch 14 and training was terminated at epoch 64. Base learning rate of 1e-2 was decreased to 1e-3 at epoch 34, and 1e-4 at epoch 54, but this learning rate reduction did not improve model results.*

Model results can be compared to true values through a scatter plot, where the x-axis represents the predicted value and the y-axis represents the true value (Figure 16). A perfect fitting model would results in all predictions made on the black dashed line y = x. Figure 16 highlights that the majority of results are clustered around the true prediction line, however the outliers of the model tend to be above the dashed line. This implies the model does not seem to be able to handle predicting very soft tissue that has a high tissue compliance, as the model never predicts a compliance greater than 1300 mm/MPa. Choosing a different loss function, such as mean squared error, which has a greater penalty on outliers, could reduce these errors but may come at the expense of absolute error or absolute percent error.

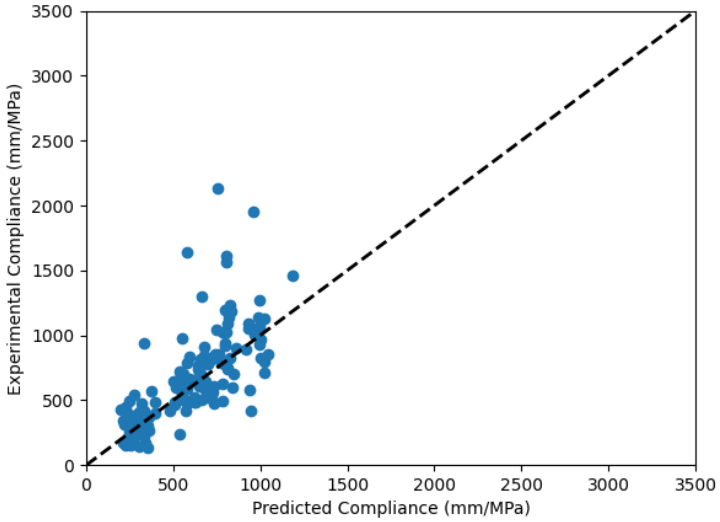

Figure 16: Demographic network predictions vs true values. *The dashed black line shows a model with 100% fit.*

The distribution of absolute percent errors is shown in Figure 17. Only several predictions were over 70% percent error, and the majority of test data results are right skewed, where most data points are contained in the histogram bins at the left end of the plot. The model predicts most results accurately, but struggles with a few outliers.

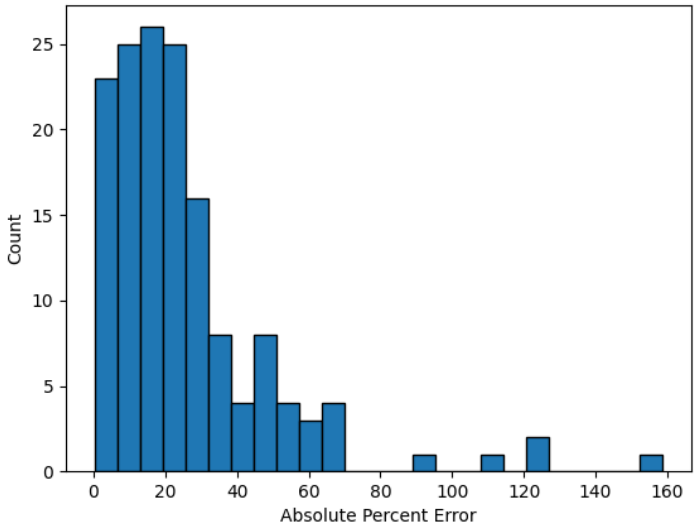

Figure 17: Absolute percent error histogram for demographics network.

An interface (Figure 18) was generated for this demographics model using the Python library Gradio (Abid et al., 2019). This is a representation of how the machine learning model could be used in the field by the novice, i.e., a translational setting (although as discussed in Chapter II, the underlying data is not supportive of this usage). Deployment of a machine learning model is an important consideration, as it is important for users to be able to use a model. This model possesses the unique benefit in that it requires no processing for results to be generated: new predictions can be made on the fly. Ultrasound images are not necessary for this model to make predictions; this model relies only on basic sign-up sheet information that can be readily available to a caregiver when a patient visits. This is useful as it requires no radiologist expertise like the models in Chapters III and V.

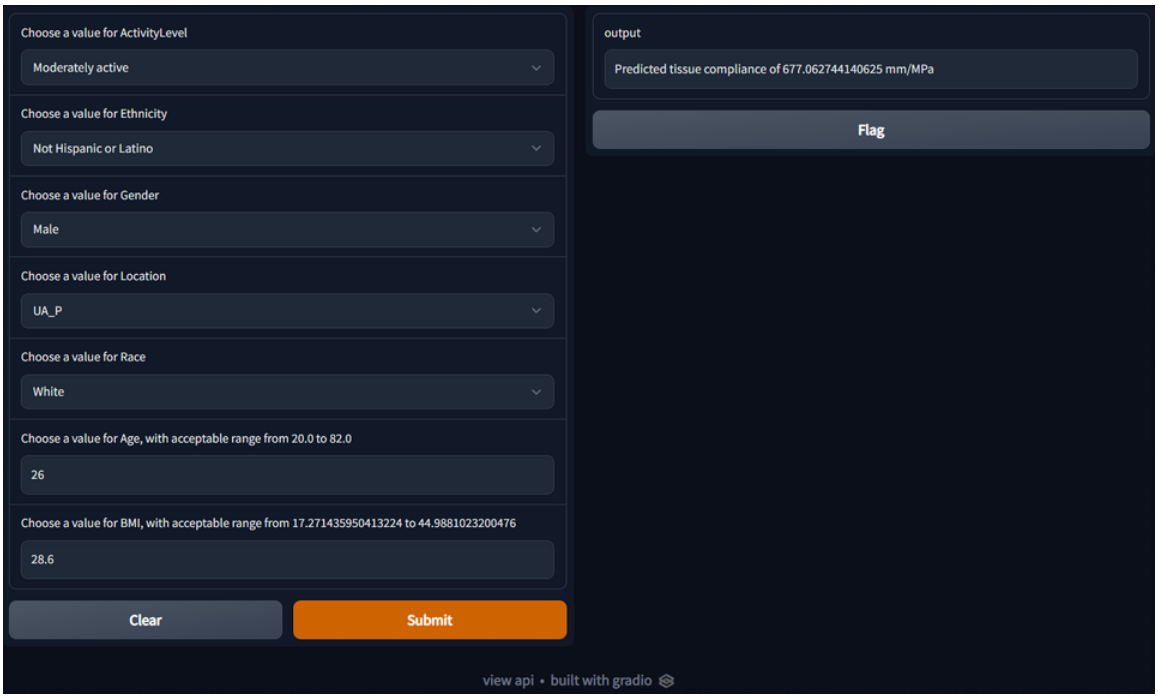

Figure 18: Interface for user prediction of tissue compliance. *Categorical values such as activity level are selected with a drop down menu, while numeric values are input from the user.*

A common criticism of machine learning models is that they operate as black box models, where the individuals who build and use the models do not understand how inputs affect the model. This is somewhat true, as the model's weights and biases are largely incomprehensible to individuals when compared to a simple linear fit's slope and intercept. Methods have been developed to gain insight into what input features most strongly influence a model's predictions. One of the most popular methods used is based on game-theory and is called shapley additive explanations (SHAP) (Lundberg and Lee, 2017). The SHAP values that are produced provide insight into why a model predicts a value outside of the mean. Figure 19 highlights the SHAP values for a specific research subject, whose values were listed in Table 8. The plot contains only the top contributing features, which are all regions. This suggests that region is a key indicator for prediction of tissue compliance, and other demographic information may play only a lesser role.

Figure 19: SHAP force plot. *Every region is listed, as the SHAP values are calculated on the one-hot encoded data. Only values that contributed 5% or more of the prediction are labeled.*

SHAP values can also be evaluated over the entire test dataset, which can give a better picture of how each input feature influences the model. SHAP values are calculated on the one-hot encoded data, hence why all categorical data is present in Figure 20. Figure 20 shows each category or value directly influenced the model's prediction of tissue compliance. Features in Figure 20 are sorted by the mean absolute SHAP values,

where higher values are indicative of greater importance. It can be seen that region is the clear most important factor in predicting stiffness.

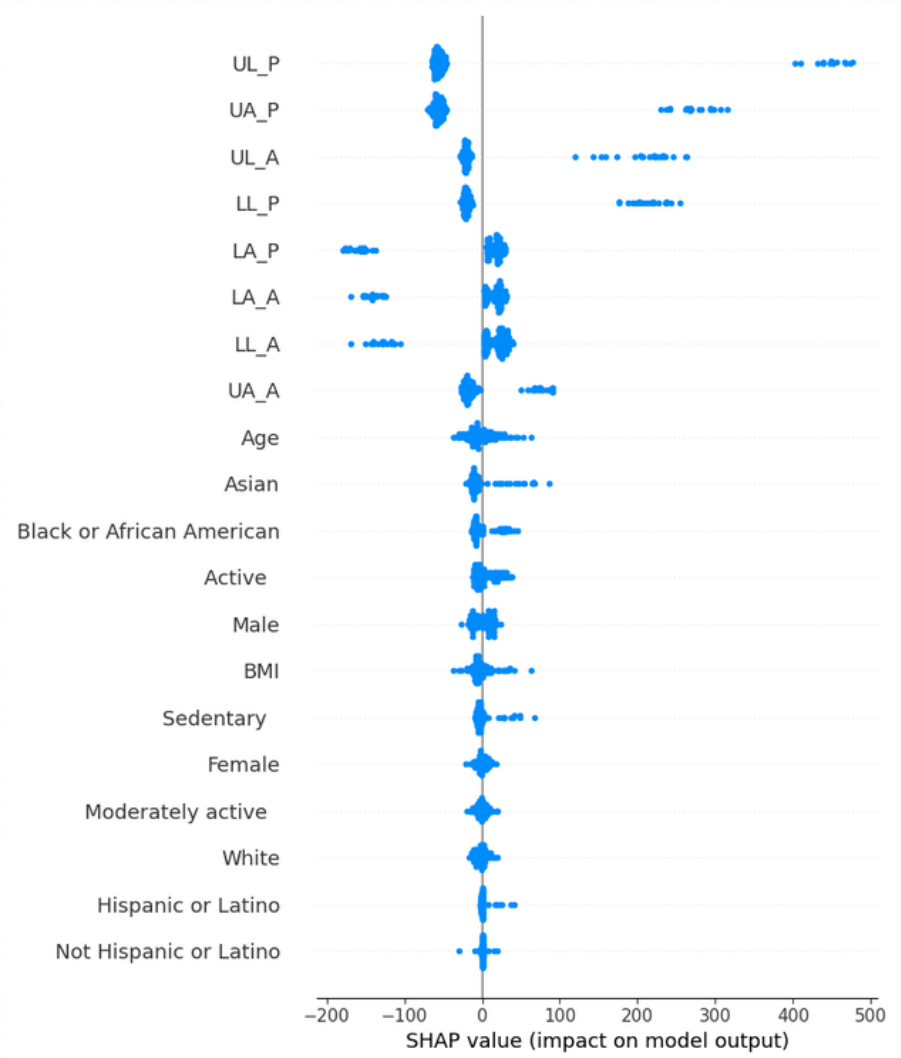

Figure 20: SHAP values for each prediction on the test data. *Note that values are sorted by mean absolute SHAP value from the one-hot encoded data, which leads to all regions having a value for each data sample (1 for the region indented indicating true, 0 for all other regions indicating these regions were not indented).*

Other demographic information can have some influence on prediction, shifting model predictions by as much as 100 mm/MPa. However, the mean absolute SHAP value is below 20 for all non-region variables (Figure 21). This type of plot could be used for feature selection if further model refinement is desired, as Figure 21 shows that ethnicity

has minimal effect on this model. Note that the minimal effect of ethnicity may stem from the lack of representation of this group in the dataset, as described in Chapter II. Figure 20 and 21 were generated on the test dataset, where there were only 8 Hispanic or Latino individuals, making distinguishing trends on this factor even more difficult.

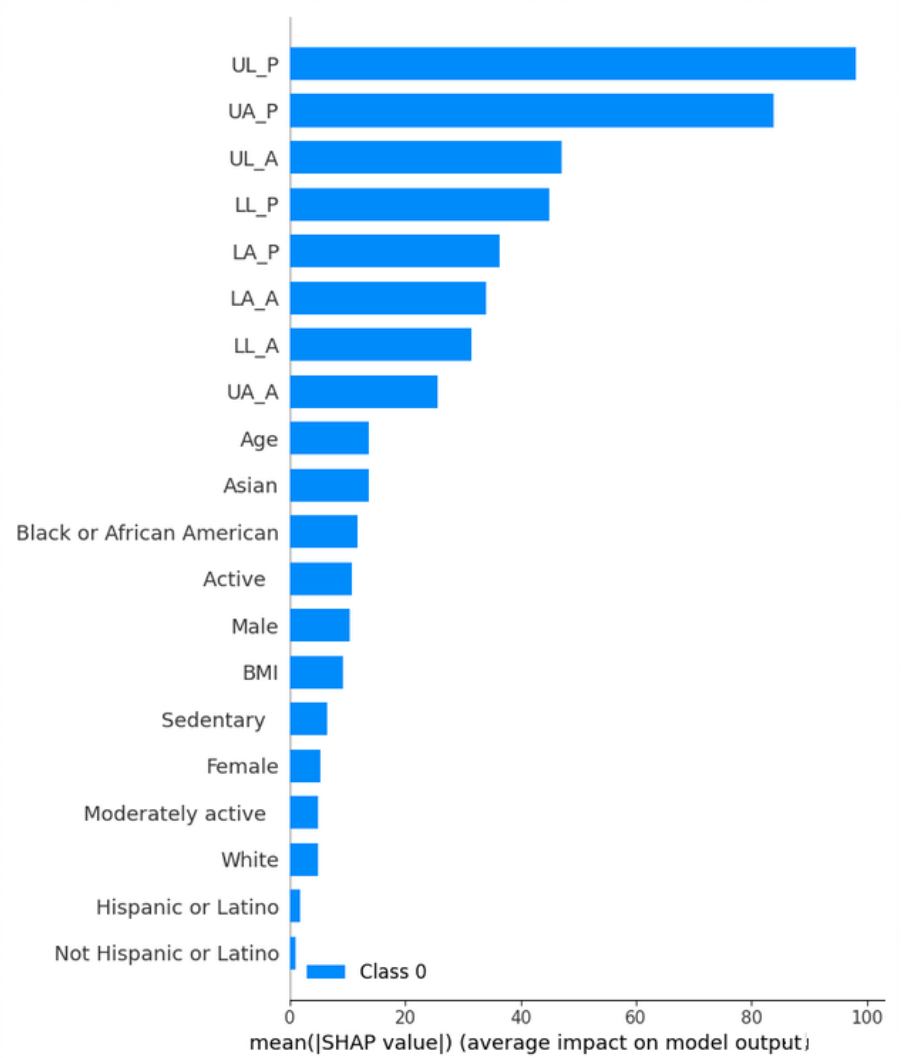

Figure 21: Mean absolute SHAP value for the demographics network. *Higher values indicate higher impact on prediction of tissue compliance. Magnitude indicates impact on prediction.*

A second neural network was formulated with a very similar setup and architecture to the demographics-based network. This model had two additional features supplied to the model: fat layer thickness and muscle layer thickness. Skin thickness was

considered for the model but was excluded since the inability to capture the small variability in skin thickness means the measurements made likely have minimal significance to predictions. Any significance that would be derived may just result from fitting to the noise inherent in measuring the thickness of the skin layer. These features were considered to be a valuable input, as they are markers from the ultrasound image that may be predictive of mechanics. It was hypothesized based on the work in Chapter III that the layers of tissue would respond similar to springs under loading. These features were assumed to be a primary driver of mechanics and Figure 8 shows a weak linear correlation between both variables and tissue compliance. It was therefore assumed that fat and muscle thickness would be a valuable input to any statistical learning model.

As described above, the model generally mirrors Figure 4, with an input normalization layer (now with two additional features in fat and muscle thickness), two hidden layers, and a single neuron for the output layer. Hyperparameters were tuned over 20 iterations using a random search of the average of two trials at each iteration. The hyperparameters consisted of the same potential values and search methods as Table 9. The optimal learning rate for the thickness+demographics model was 1e-2. The best number of hidden neurons was 160. For the three hyperparameters designed to reduce overfitting, gaussian noise was not included, a dropout rate of 0.2 was used, and an L2 regularization constant of 4e-4. These hyperparameters led to very quick convergence in

the model, as the model reached its lowest validation loss at the 4th epoch, and training was terminated at the 54th epoch (Figure 22).

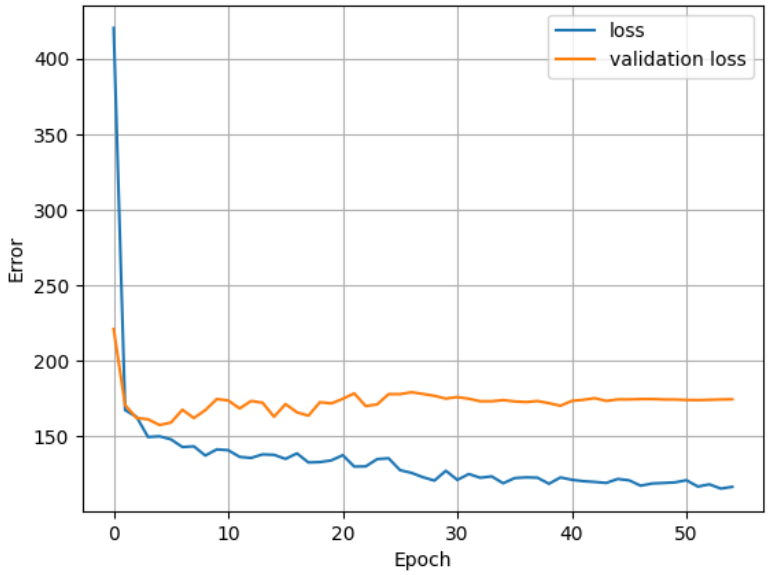

Figure 22: Training history for the thickness+demographics network. *Model converged at epoch 4, and then began to exhibit overfitting.*

Model results were surprising and ran against the hypothesis that tissue thicknesses would improve model performance. While this model produced a lower MAE at 157.17 (compared to the demographics-based model MAE of 158.42), MAPE increased to 26.33 (compared to 25.52). It is surprising at a glance to see a lower absolute error, but a higher absolute percent error. A lower absolute error, but a higher absolute percent error indicates that this model is likely performing slightly better on higher tissue compliance values, while now performing worse on lower compliance samples. For example, an estimate of 1100 mm/MPa for a 1000 mm/MPa sample results in an MAE of 100 mm/MPa and an MAPE of 10%. Meanwhile an estimate of 575 mm/MPa for a 500 mm/MPa sample results in a lower MAE of 75 mm/MPa, but a higher MAPE 15%. These

results are not especially noticeable in Figure 23, as the difference between the two model's performances is likely close to negligible.

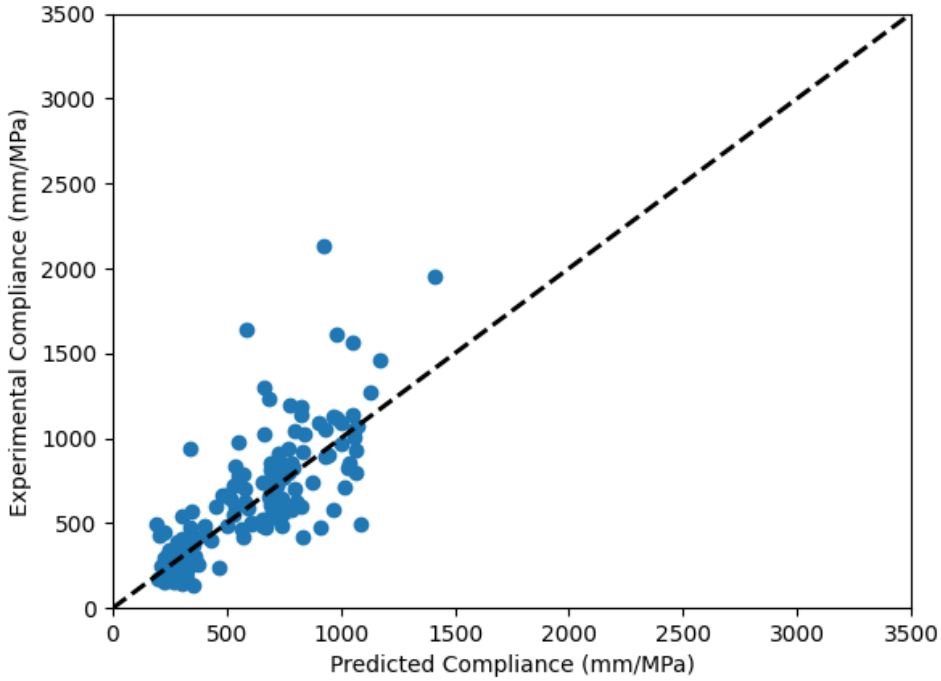

Figure 23: Thickness+demographic network predictions vs experimental compliance. *A perfect model would fit to the black dashed line.*

The SHAP values for the thickness+demographics model provided some interesting results. A force plot was again generated for the research subject in Table 8, but with the two additional measurements of fat thickness and muscle thickness (Figure 24). The subject had a fat thickness of 5.477 mm, and a muscle thickness of 20.369 mm. Muscle thickness was the second most important feature for this specific subject, and BMI was now considered an important feature when this value was previously not important in Figure 19.

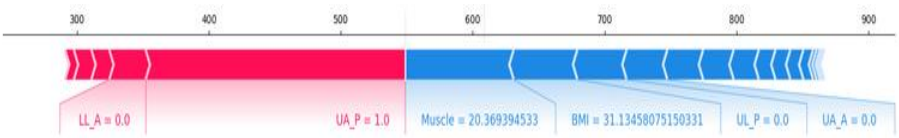

Figure 24: Thickness+demographics SHAP force plot. *Only the features with over 5% contribution have their names shown.*

Figure 20 and 21 showed that in the demographics-based model, consideration of each region made the top 8 SHAP values. This was not the case in the thickness+demographics model, where both muscle and fat thickness were in the top 8 values, as the 2nd and 4th values respectively (Figure 25). Additionally, BMI became the 8th most important SHAP feature, whereas it was the 14th most important feature by SHAP value in Figure 21. While the new inputs of muscle and fat thickness are in the top 4 most important feature of the model, the lack of model improvement highlights some of the challenges in interpreting machine learning model decisions.

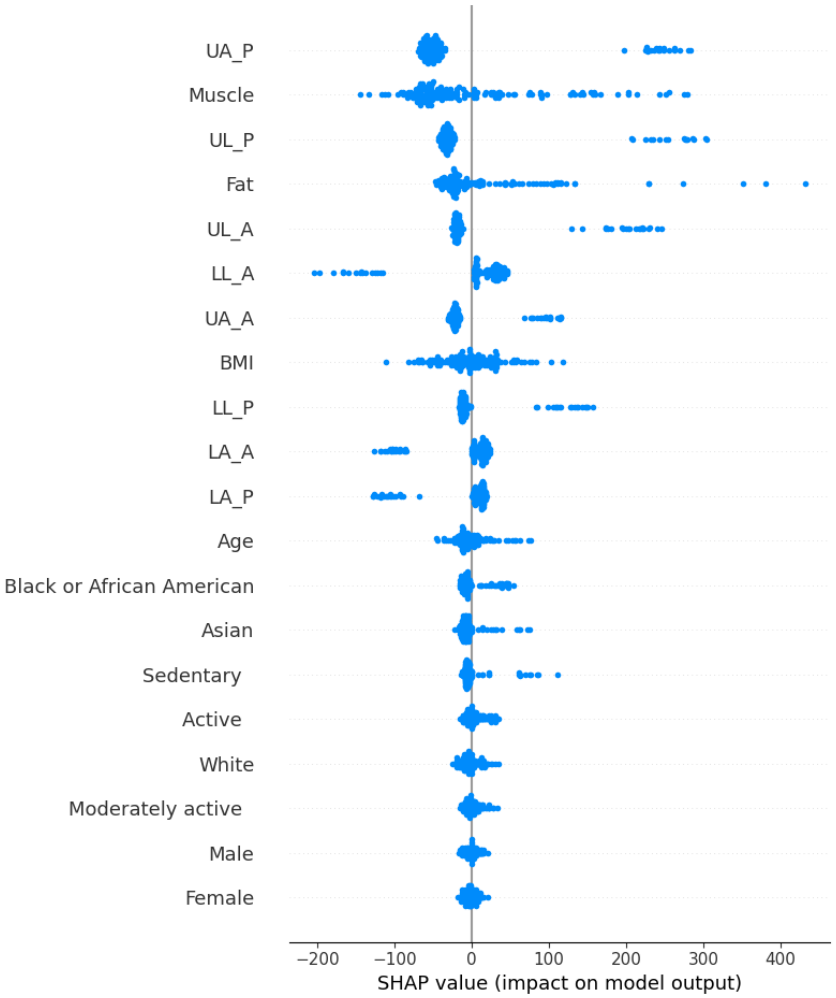

Figure 25: Thickness+demographics prediction SHAP values. *Note that values are sorted by mean absolute SHAP value from the one-hot encoded data, which leads to all regions having a value for each data sample.*

While fat and muscle thickness played an important role in prediction of surface compliance values, model accuracy did not increase. Some of this may stem from the fact that muscle and fat thickness likely has a relationship with region and BMI. Figure 8 showed a weak positive trend between fat and muscle thickness and BMI. It is expected that for most of the healthy population, the upper leg will have more muscle and fat than the lower arm (Figure 26). Mean muscle thickness and mean fat thickness both fit reasonably well to compliance data when grouped by region, suggesting some fat and muscle thickness information was already accounted for in the demographics-based

model. Given this, differences in tissue compliance as a function of layer thickness may have already been factored into the region designation.

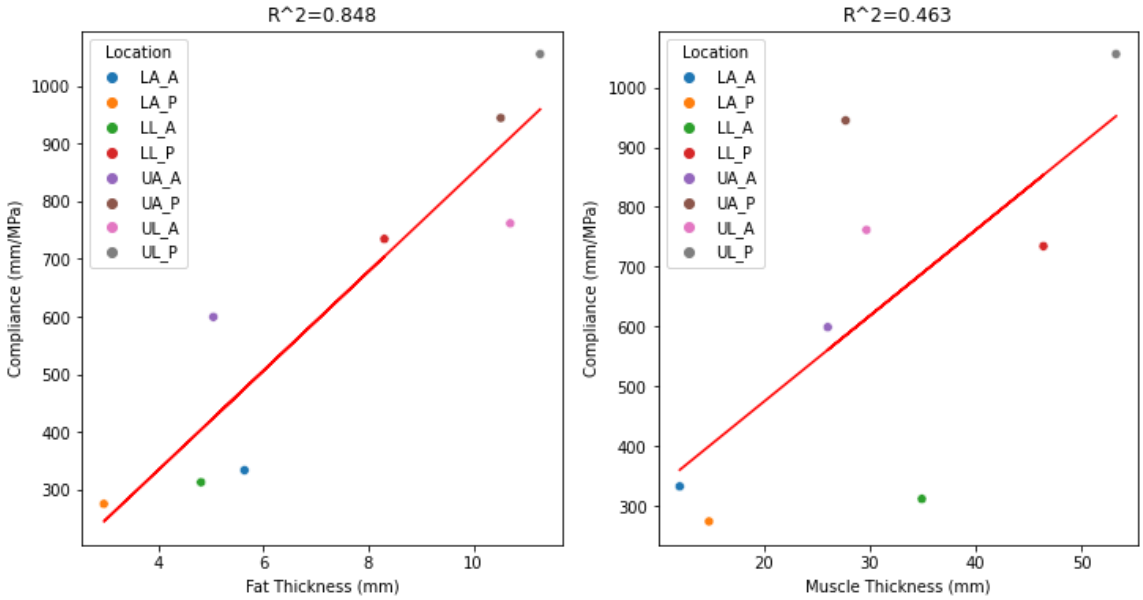

Figure 26: Mean muscle and fat thickness by indentation region. *Red line is a linear regression fit. Note linear fit is shown just for an example, machine learning models will fit with higher complexity equations when appropriate.*

Discriminative ability for the prediction of compliance from the addition of fat and muscle thickness would not stem from the grouping of region, but instead within region differences. Shown for the upper leg anterior region, it can be seen that the fits become essentially random when selecting a single region (Figure 27).

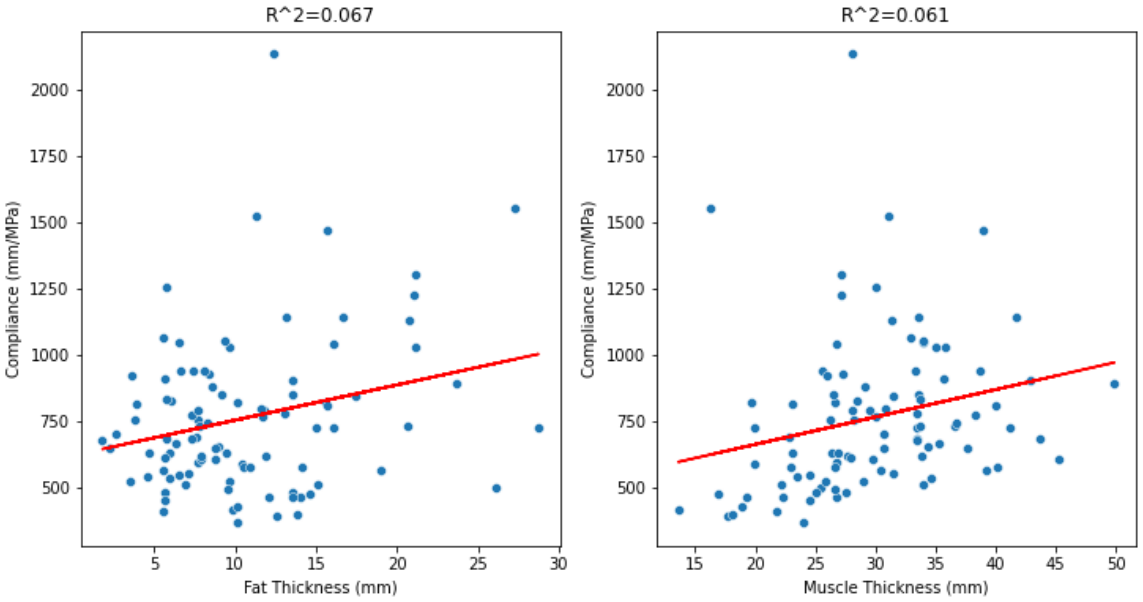

Figure 27: Muscle and fat thickness fits to compliance for anterior upper leg. *Predictive ability of fat and muscle thickness becomes weak once controlling for region. Note linear fit is shown just for an example, machine learning models will fit with higher complexity equations when appropriate.*

For indentation at the same region, subjects with a greater amount of fat or muscle (which would yield to higher thicknesses) likely have a higher BMI. This relationship between BMI and tissue thickness likely leads to BMI's higher importance in this model, whereas in the demographics-based models there is no numeric difference between regions for BMI to have much influence (the only non-categorical variables in the demographics-based model was age, which is also constant across regions, while fat and muscle thickness vary across regions).

Both the demographics network and the thickness+demographics network have error that is higher than would be desirable, but the results are not necessarily poor. As elaborated upon within Chapter II, there are limitations in the dataset itself. It is unclear with the current dataset if 25% reaches the limit of experimental variability, or if this performance is undesirable. The current dataset does not provide the ability to test this important factor. Results are reported as a float value, but the accuracy in which a

difference in compliance is noticeable in a surgical simulator or significant for a compression garment may have large bounds. Can a surgeon even notice a difference of 100 mm/MPa surface compliance? Classifying each tissue into bins defined by stiffness ranges could “improve” model performance, by providing a range for each prediction to fall into. Machine learning is commonly used on classification tasks. The bins could have ranges that are similar to the bounds in which the change in haptics from different compliances is noticeable. Given most test data points are near the line of perfect fit for the model (Figure 16, Figure 23), these deep neural networks may be sufficient at classifying tissue compliance.

It is difficult to justify the manual annotation of ultrasound images for layer thicknesses when these values possibly only serve as a proxy for region of indentation. While a machine learning model can fit a more complex fit than a linear fit, the distribution of points in Figure 27 shows that there is likely no correlation between the layer thickness and compliance after region is controlled for. The relationship between tissue thickness and other variables besides BMI may be noteworthy however.

The deep learning models shown performed favorably compared to the linear mixed effect models described in Chapter III. While technically training on fewer data samples (as these machine learning models needed validation data), the models had very similar mean absolute percent errors. The demographics network had the lowest mean absolute percent error of any of the models yet. The demographics model has the added benefit of not requiring any medical equipment or radiologist interpretation of layer thickness when compared to the demographics+thickness model and the statistics models of Chapter III.

CHAPTER V

CONVOLUTIONAL NEURAL NETWORKS FOR PREDICTION OF INDENTATION RESPONSE DIRECTLY FROM ULTRASOUND IMAGE

Chapter V introduces the convolutional neural network, the most complex form of modeling in this thesis. Convolutional neural networks are built upon the foundational neural networks of Chapter IV, in addition to the mathematical operation of convolution. Mathematically, convolution is used to relate one function to another. It can be used to show how much one function overlaps with another (Weisstein, n.d.). This can be extended to two matrices, which is where the power of a convolutional matrix lies. A matrix can be compared to another matrix, which is typically much smaller (and typically known as a filter), to extract features from that original matrix (Figure 28). When one considers the many forms of data that can be represented as a matrix, such as text, images, or video, the convolutional neural network becomes a powerful tool. Filters can be used for feature extraction, such as edge detection, image sharpening or blurring, although it is up to the model to decide which filters to use (Albawi et al., 2017).

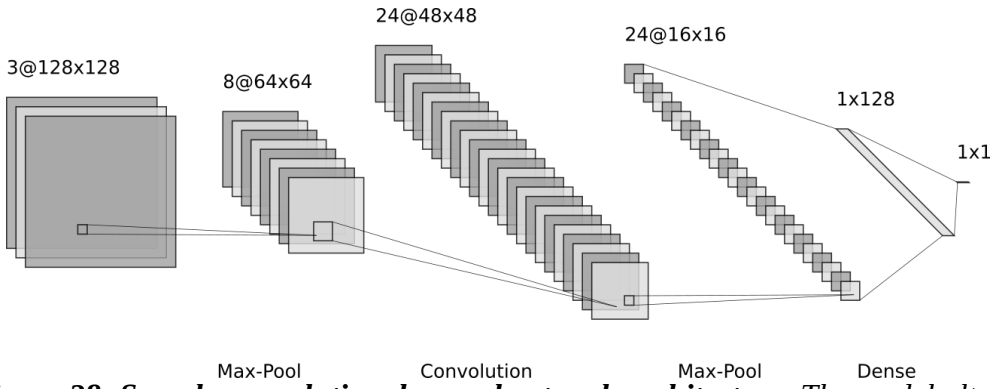

Figure 28: Sample convolutional neural network architecture. *The model alternates between convolution and max pooling layers, before adding dense layers for the prediction of a single value.*

Convolutional neural networks began to rapidly spread in the 2010s, as these networks dominated computer vision tasks such as the ImageNet Large Scale Visual Recognition Challenge (ILSVRC). This challenge involved classifying different images into what the focus of the image was, but the difficulty of the task was there were 1000 categories in 1.2 million images (eventually expanded to 22,000 categories in 15 million images) to predict (Russakovsky et al., 2015). In 2012, Alex Krizhevsky introduced "AlexNet" a convolutional network that dominated the competition (beating second place by over 10%), based on stacking of convolutional layers (Krizhevsky et al., 2017). The network also leveraged the concept of dropout to prevent model overfitting. The remaining ILSVRCs were all similarly won with convolutional neural networks, and architectures increased in depth as GPU computing power increased. Some of the most popular in image classification include AlexNet, ResNet50 (He et al., 2016), Xception (Chollet, 2017), and VGG19 (Simonyan and Zisserman, 2014). Other architectures are available for different categories of tasks, such as medical image segmentation (Ronneberger et al., 2015), or video analysis (Hara et al., 2017).

The above discussion on network architecture was provided to highlight that network architecture has become quite complex and specialized by task. It is not necessary (or feasible) for most researchers to develop their own computer vision convolutional neural network, as they have become immensely complex. While it was relatively simple to develop a multilayer perceptron network in Chapter IV, this chapter will use an “off the shelf model” to skip much of the model architecture development since using ultrasound images turns this problem into a computer vision task.

Two separate models were developed using convolutional neural network methods. The first model relied only on the ultrasound images alone, without any demographic information. The second model was given the ultrasound image, demographic information, and also some of the image metadata.

Focusing on the first model (images only) to start, the dataset again consisted of 752 datapoints, except the datapoints were now ultrasound images. The images were pulled from multisbeta.edu by using “query data” to access accepted AND “Analysis and Visualizations” with “Ultrasound minFrame". The minFrame condition pulls the ultrasound data associated with the initial state of indentation, which is of interest for this work as it contains the image of the region at its unloaded anatomical state. The 752 images were in DICOM (.ima) format, which is unsuitable for neural network training in Tensorflow natively. Images were converted to portable network graphic files (.png). Images were also processed in an effort to remove extraneous information from the image that the neural network might erroneously focus on, including text, the orientation marker, and pulse repetition line. This task was automated with the Python library opencv (Figure 29). Images were cropped to a size of 676 x 676 pixels.

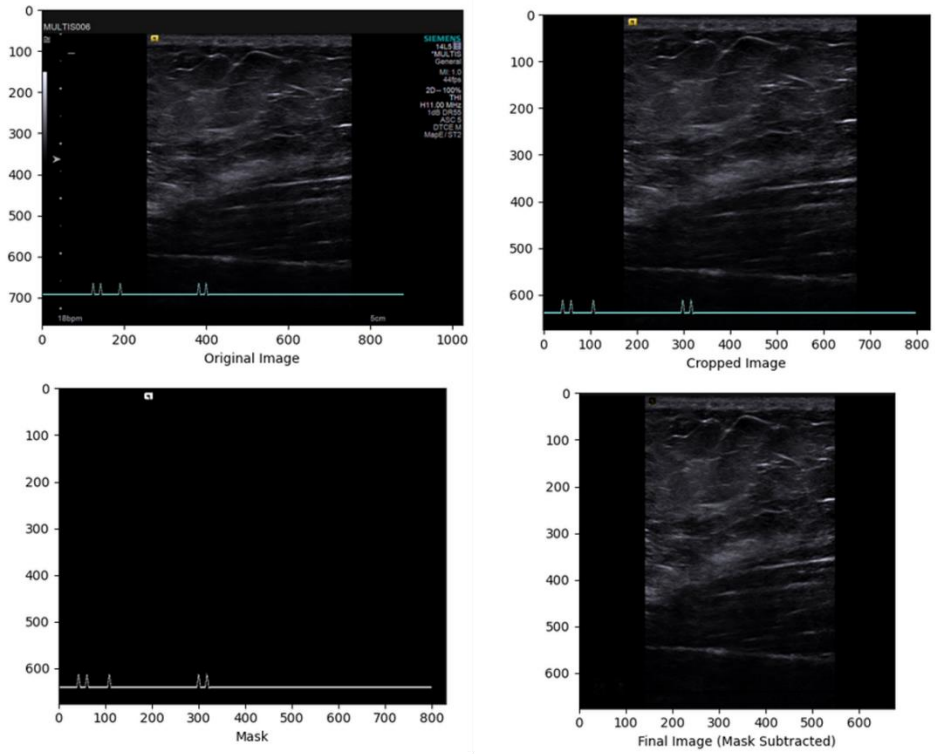

Figure 29: Preprocessing of ultrasound images. *(Top, left) Original image. (Top, right) Cropped image. (Bottom, left) Mask to be subtracted from the cropped image. (Bottom, right) Final processed image.*

With the processed images, another key feature of image processing workflows in machine learning models was implemented, with data augmentation. As with the models in Chapter IV, overfitting is a concern for convolutional neural networks. Models such as Xception can have over 200 times as many parameters (where parameters is defined as any weight or bias in the model) as the simpler models of Chapter IV, totaling over 20 million parameters. VGG19 is even large, at over 143 million parameters. This many parameters can lead to model performance stagnating quickly as the model essentially memorizes each input image. One of the most effective methods to combat this is to guarantee the model never sees the same image more than once. With data augmentation, several random different translations and modifications are made to the image every time the image goes through the model (in fact, the model never actually trains on the original

image, but rather only augmented versions of it). The list of data augmentations performed can be found in Table 10. These augmentations are only performed on the training data set, as the validation and test datasets are kept in their original form. The lack of augmentation to the validation and test datasets is due to the fact the model does not train on these images and preserving the original images allows for tuning and evaluating model performance on actual real-world images.

Table 10: Data augmentation parameters for ultrasound images. *Augmentation value ranges were taken from literature (Heo et al., 2019). Values for each parameter are chosen randomly for each image each iteration through the training dataset.*

Augmentation	Potential Range
Preprocess Image Values	[-1, 1]
Horizontal Flipping	0 or 1 (No or Yes)
Vertical Flipping	0 or 1 (No or Yes)
Width Shift Range	[-0.2, 0.2]
Height Shift Range	[-0.2, 0.2]
Rotation Range	[-30, 30]

Sample data augmentations applied to one image with its associated compliance are shown in Figure 30. Note that preprocessing image values has no real effect on the image. The Xception machine learning model that was used requires data in the range -1 to 1, but images can easily be scaled back into 0 to 255 for visualization. Width shift and height shift are similar, where the image can be shifted left or right up to 20%, and up or down 20%. New pixels are colored based on the nearest pixel in the original image. Rotation will rotate the image 30 degrees in either clockwise or counterclockwise direction. The value of some augmentations is clear: a horizontal flip is a valuable

transformation that essentially doubles the dataset size, as it corresponds to rotating the ultrasound probe 180 degrees. Some augmentations are more valuable for a machine, such as the vertical flip, as viewing the image upside down is not intuitive for humans. The information in the image is the same though when flipped upside down, and techniques such as this help machine generalize.

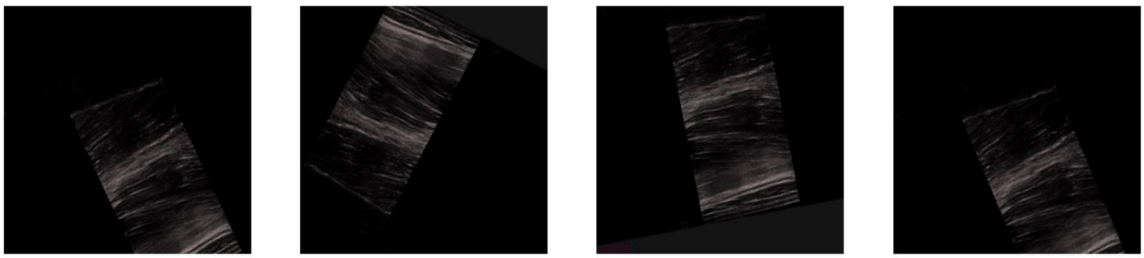

Figure 30: Data augmentation applied to a single image. *The ultrasound image is taken of the posterior upper leg.*

With the images ready for input, model selection was the next important step. Ideally, several different architectures should be experimented with, but computational resources were a limiting factor with some of these convolutional neural networks. For instance VGG19 is nearly six times larger than Xception (Chollet, 2017). Additionally model comparison in medical imaging application like lung segmentation can yield small percent differences under 4% when comparing different off the shelf networks (Heo et al., 2019), a negligible difference for exploratory research. As such, computational resources were prioritized and only Xception was used due to its relatively small size. Future work should consider expanding model testing.

The base model of Xception consists of 36 different convolution layers with model architecture set up to separate spatial correlations from cross channel correlations, where cross channel correlations are differences between the red, green, and blue channels (Chollet, 2017). Since the dense layers on top of the convolutional layers in

Xception were used for image classification rather than a regression problem, only the base convolutional layers were used. A “GlobalAveragePooling2D” layer aggregates features from the convolutional layers to a smaller, uniform dimension of outputs (Lin et al., 2014). Similar to Figure 28, dense layers were added on top of this model. This final top portion of the model consisted of a dense layer, a PReLU activation function, and a dropout layer, before being sent the final single neuron for prediction. The base model was set to false during training, meaning only this top dense layer was trainable. Xception weights were loaded for the base model, but the dense layer has no associated weights. Given the model prepopulates the dense layer with random weights, it is important to tune this dense layer. The base layer was trained during a later step.

A hyperparameter search was performed similar to in Chapter IV. A smaller search space was used given the lengthier nature of training on images (around 25 images per iteration) While the full hyperparameter search took around 40 minutes in Chapter IV, a single iteration took 25 minutes with the convolutional neural network. Search space ranges and optimal values are shown in Table 11. The search consisted of a random search of 5 iterations with only one execution per trial. After optimal hyperparameters were found, the model was retrained with these parameters. Other model parameters include the batch size, which was set to 4 given GPU memory constraints. The model was given up to 50 epochs, although this value was never hit. The model stopped training after five consecutive epochs of no validation loss improvement. The learning rate was multiplied by a factor of 0.1 after 2 consecutive epochs of no validation loss improvement. The model compilation was identical to Chapter IV, using the Adam optimizer and mean absolute error as the loss function.

Table 11: Convolutional neural network hyperparameter search. *Searching of values is conducted with a random search of parameters.*

Hyperparameter	Potential Values	Search method	Optimal value
Learning Rate	[1e-2, 1e-3]	Choice of value	1e-2
Number of Neurons per Hidden Layer	[32, 64, 128, 256, 512]	Choice of value	32
Dropout Rate	min = 0.1, max = 0.6, step size = 0.1	Sampling at step size intervals	0.5

Model training stagnated at epoch 24, leading to training termination at epoch 29 (Figure 31). Model learning rate began at 1e-2 and was decreased by a factor of 0.1 at epochs 13, 23, 27, and 29 (terminating at 1e-6 learning rate). When making predictions on the test data, the model returned a mean absolute error of 224.44 mm/MPa and mean absolute percent error of 31.58% when predicting on the test dataset, the worst model of this work so far by around 2 absolute percent error. One aspect of the learning curve in Figure 31 worth discussing is the fact that the validation data had a smaller mean absolute error at most of the epochs. This could stem from the data augmentation used. The validation dataset had no augmentations, while the training set utilized data augmentation values from literature (Heo et al., 2019). These values may have been appropriate for segmenting medical images, but not as appropriate on ultrasound indentation mechanics. Future work may consider trying new augmentation parameters.

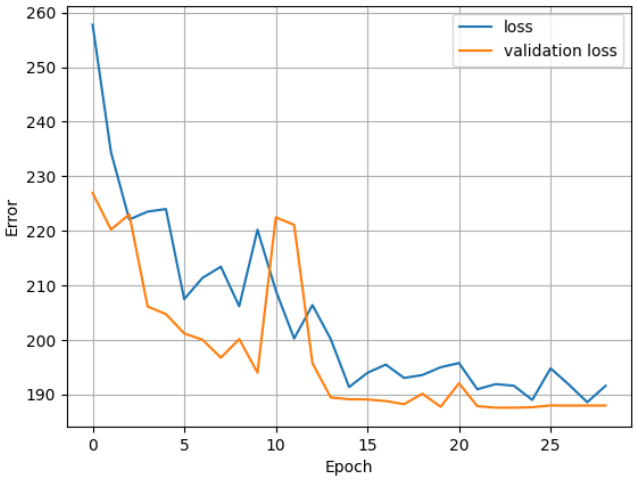

Figure 31: Image model training history. *Learning rate was decreased at epoch 13, 23, 27 and 29.*

While the deep neural network had most values contained within the first four bins of error, the image convolutional neural network has a much more even distribution up to around 50% (Figure 32). The distribution of the error shows this convolutional neural network was not able to make as accurate of predictions as the demographics deep neural network (Figure 17). Figure 32 does not share the right skew that Figure 17 exhibits, suggesting worse ability to accurately predict compliance.

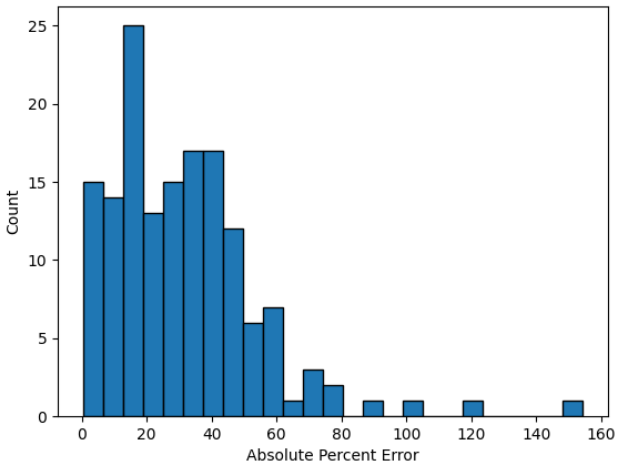

Figure 32: Image model absolute percent error histogram.

The convolutional neural network did similarly seem to struggle with some outliers, as there were 3 samples with an absolute percent error over 100%. It can be observed in Figure 33 that similar to the deep learning models of Chapter IV, this model struggled to predict many of the higher experimental compliances. The model made all predictions below 1000 mm/MPa, when the test data included subjects where the compliance was in excess of 2000 mm/MPa. This suggests that for highly compliant tissue, there may be an underlying variable that is not captured in the model's input data. This could stem from a variable like loading rate, which would affect soft tissue stiffness.

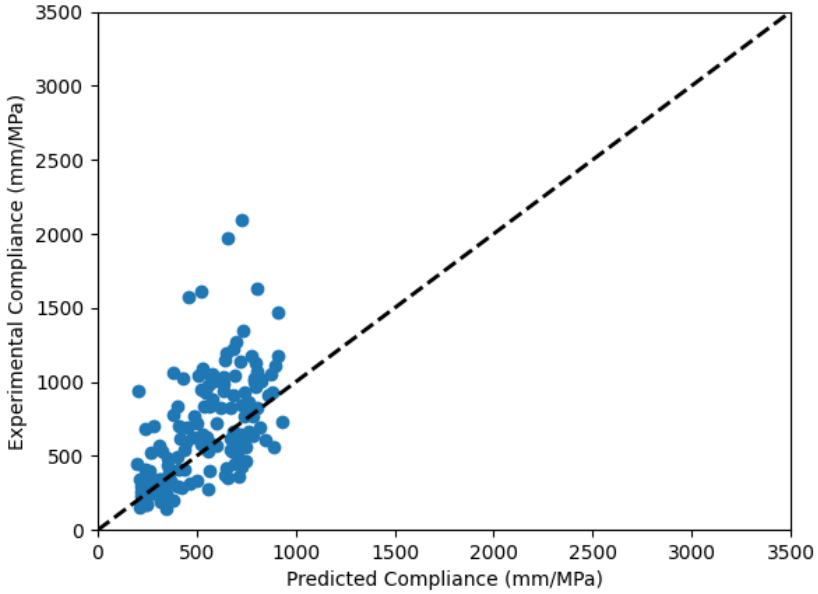

Figure 33: Image model predicted vs experimental compliance. *The black dashed line represents a model with a perfect fit to the experimental data.*

After finding optimal hyperparameters and training, the model was fine-tuned. Fine-tuning is a process of retraining all layers of a model, including the base layers that were not previously trained, with a low learning rate to better tune a model to a specific application (Bonaccorso, 2020). The small learning rate is required so the model does not overshoot the minimum. The model was set with a 1e-5 learning rate. Model training

took 11 epochs, as learning stagnated at epoch 6. Interestingly, model MAE increased slightly to 235.52 mm/MPa, while MAPE increased to 33.29%. This suggests the base layers of the Xception model were already reasonably well trained and the additional training was unnecessary. The base layers of Xception were trained on image classification of objects like dogs, making it a surprising conclusion that the base layers seemed to have already been fairly well tuned for a regression problem in the medical domain. The model may have already been at close to the maximum performance, and the additional training worsened generalization by overfitting the model.

The final model was built upon the above convolutional neural network by considering how a model would perform if it had the image to learn from, in addition to the demographic information associated with each image. Heo et al. showed that incorporating demographic information into a machine learning model can improve model performance (Heo et al., 2019), and the models of Chapter IV outperformed the image only model. Trends and insights from the image may supplement the demographic deep neural network predictions, which had a higher mean absolute percent error than any of the linear models, or either of the machine learning models explained earlier. Ultrasound image metadata was also a factor in the model, to provide some more image based information for the model to use. Within the DICOM metadata, the Physical Delta X, representing the physical dimensions for the region shown. Physical Delta Y was identical to Physical Delta X in all images and is a single float value, providing an avenue for informing the model about the spatial dimensions in the image by adding this information in with the demographics.

Model setup was largely similar to the image convolutional neural network, utilizing the base Xception model. The Xception model's layers were set to untrainable at first, before being fine-tuned after model training. Ultrasound images were preprocessed as in Figure 29. Images were used as an input for Xception. The top layer of the model was redesigned however, to incorporate the image metadata and demographic information. The output of the Xception model was fed to a global average pooling layer to extract important model features, followed by a dense layer of neurons, with a PReLU activation function, and a dropout layer. This layer is then concatenated with the normalized demographic and metadata. This combined layer then passes into two consecutive sets of a dense layer, PReLU layer, and dropout layer, so the model can process the new inputs, before moving to the prediction layer (a single neuron).

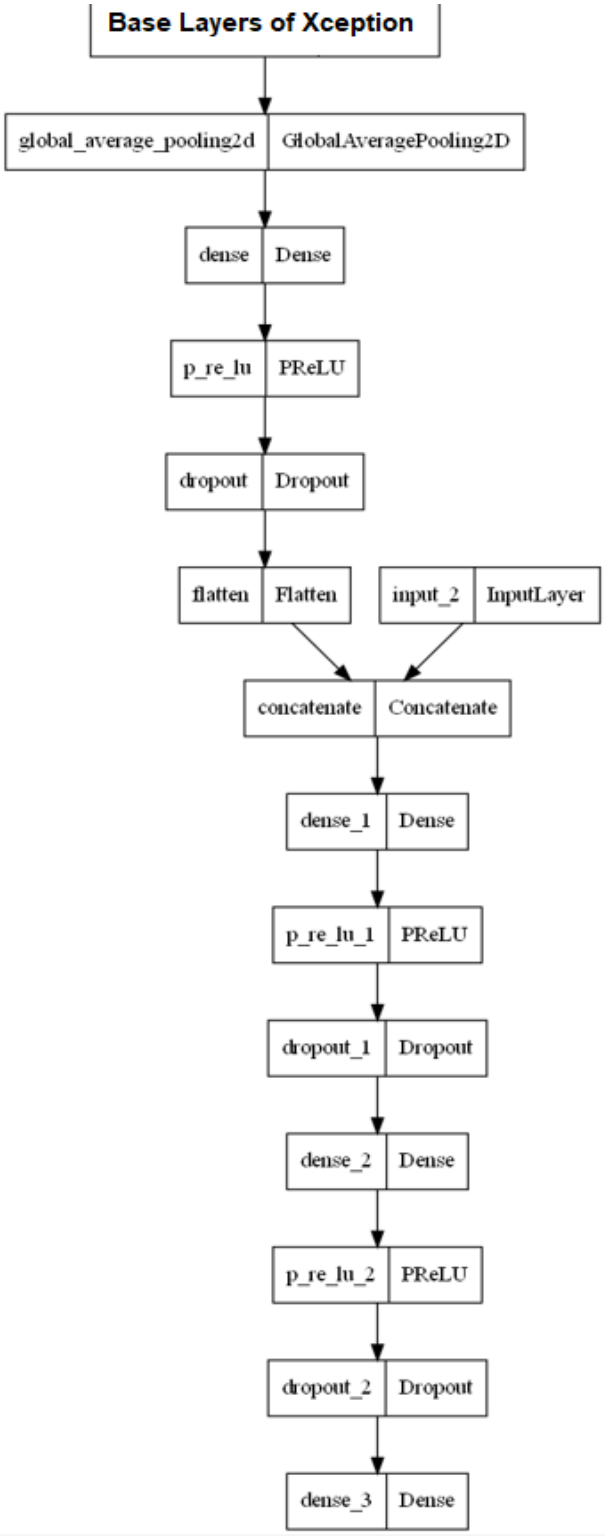

Figure 34: Overview of image and demographics model architecture. *The Xception layers were simplified to a single block. "input_2" represents the demographics and image metadata.*

Some of the limitations of generating a multi-input model led to a few of the features of the previous image convolutional neural network being dropped. For instance, with this model a hyperparameter search could not be performed, as KerasTuner was incompatible with a multi-input model. As such, the learning rate was set to 1e-2, the number of neurons was set to 32 (identical to the above image model hyperparameter search), and the dropout rate was set to a lower value of 0.2 (based on manual iterative testing). These values are not necessarily optimal. Batch size was kept at 4 samples. The model was given up to 100 epochs again. The model stopped training after five consecutive epochs of no validation loss improvement. The learning rate was multiplied by a factor of 0.1 after 2 consecutive epochs of no validation loss improvement. The model was compiled using the Adam optimizer and mean absolute error as the loss function.

Model training took 25 epochs. The learning rate decreased at epoch 18, 20, 23, 25. Figure 35 highlights a similar behavior as seen in Figure 31, where the training data loss exceeded the validation loss in most epochs. Modifying the image data augmentation methods would likely be a worthwhile task, as the augmentations may be removing some of the key information in the image. A better selected choice of augmentations could lead to further model training. This could involve a user incorporating these augmentations values into a hyperparameter search, or through a literature search. The augmentation parameters used were originally formulated for x-ray images, and ultrasound images may perform better under a different set of augmentations. Extra computational resources would prove beneficial to quickly explore different values.

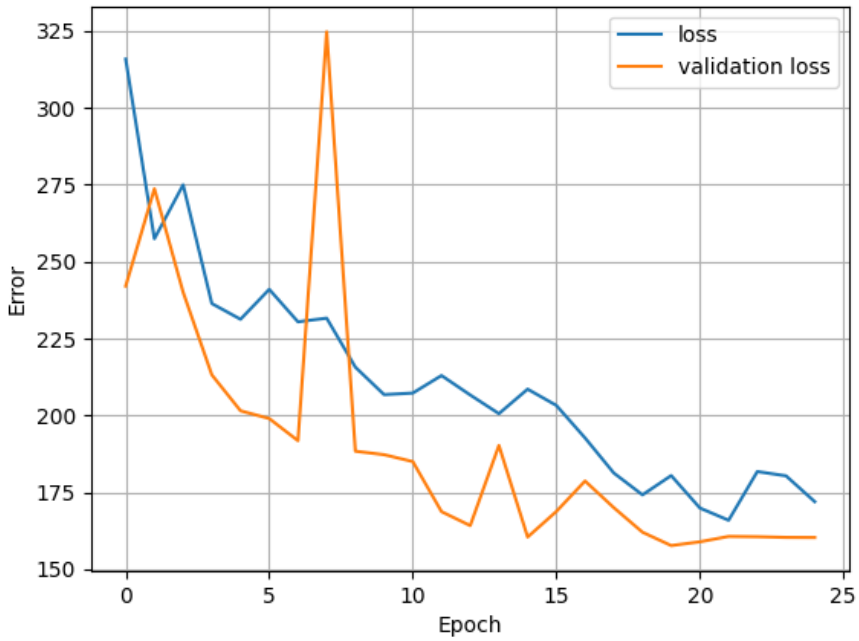

Figure 35: Image and demographics model training history.

The model achieved a mean absolute error of 183.48 mm/MPa, and a mean average percent error of 25.90%. The model is comparable to the demographics network in terms of mean absolute percent error, but underperforms in mean absolute error by 25.06 mm/MPa. As described earlier in Chapter IV, a higher mean absolute error paired with a similar mean absolute percent error is indicative of the model having poor accuracy on higher compliance data points. For instance, Figure 36 shows that the data points furthest from the black dashed line of perfect fit occur at higher experimental compliances. Model underperformance on these high compliances has been consistent across modeling techniques, suggesting modeling these high compliances may be outliers and will be difficult to predict with a data driven approach.

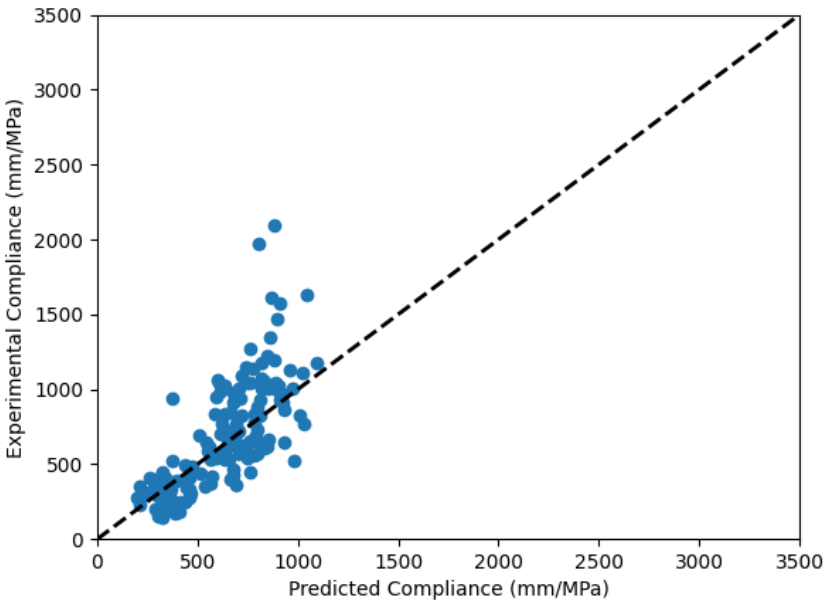

Figure 36: Image and demographics model predicted vs experimental compliance. *A perfect fitting model would fall on the dashed black line.*

Distribution of mean absolute percent error was similar to Figure 32, in that the percent error was largely distributed between the first 10 bins under around 50 absolute percent error (Figure 37). This model did outperform the image-only network by almost 8 absolute percent error, some of this was due to fewer large outliers. Figure 37 shows that this model produced only a single prediction with over a 100 percent absolute error out of the 151 test samples. This compares favorably with the demographics network, where 4 subjects exceeded this threshold and also to the image-only network which had 3 subjects exceed 100 absolute percent error. This model can then be considered to make reasonable although less accurate predictions for most subject, but is lacking in precision.

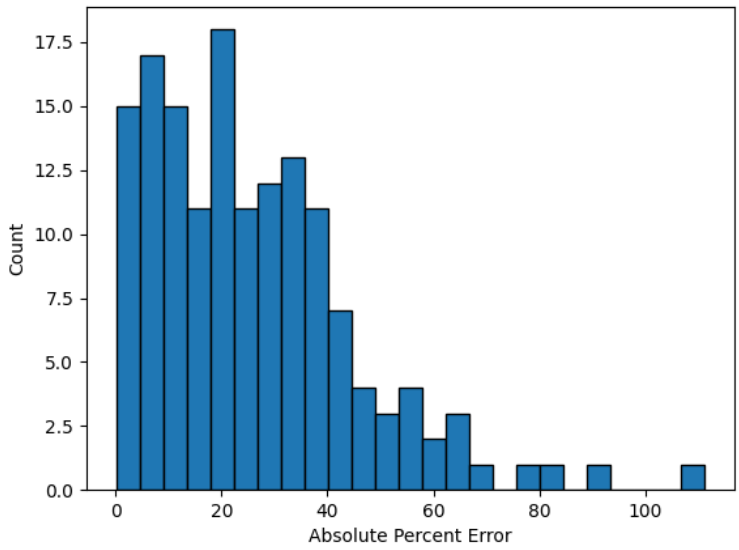

Figure 37: Image and demographics model absolute percent error histogram.

Fine-tuning was performed on the model with a small learning rate of 1e-5. Fine-tuning once again produced mixed results, slightly decreasing model mean absolute error from 183.48 to 180.02, but model mean absolute percent error increased to 28.78%. As before, it could be the base layers of Xception are already reasonably well trained or the fine-tuning process may need to be improved with a more rigorous testing of learning rates.

The most complex models of the thesis come from this Chapter, and with added complexity comes the expectation of superior performance. Ultimately, these models failed to meet those expectations, although the image and demographics model performed comparable to the best performing deep neural network and the best performing linear model based on mean absolute percent error. This model does suffer from a black box design, which makes it difficult to recommend over the previous models. While these models do not require radiologist annotation of ultrasound images, the lack of explainability make it difficult to justify the model when there are linear models

performing just as well. Black box design in the clinical realm can be a dangerous practice, as failure to understand the mechanism behind a process or treatment can lead to discovery of unintended side effects later on. The image and demographics convolutional neural network could begin to outperform the other models with additional data, but further data collection is needed to test this hypothesis. The models may have reached their performance limit, as the inherent noise in mechanical measurements may exceed 25%. If this is the case, several models are near maximum performance. Further experimental data collection would help test this theory, but for now it is difficult to say whether the models are at maximum learning capacity or if there is further room for improvement. As the work stands without significant performance gains, it is not recommended to use machine learning models over a simpler model.

CHAPTER VI

CONCLUDING REMARKS

This thesis explored the potential of data-driven models for prediction of tissue compliance during indentation of musculoskeletal extremities. Chapter I described our motivation, while introducing the models utilized in this paper. Chapter II introduced the data, a key part of data driven modeling given the data informs how the model makes decisions. Chapter III, IV, and V showed different ways the data could be used to make statistical learning models and how decisions were made for those models. Performance between models was quite similar across the Chapters (Table 12). The best performing statistical model (location-specific linear model) had a mean absolute percent error of 25.68%, the demographics network had 25.53%, while the image and demographics convolutional network was at 25.90%. Even the worst performing model, the image convolutional network, was only around 6% worse than the best model by mean absolute percent error. This suggests the models are close to interchangeable. With such small differences in model performance, the simplest model is likely recommendable in the location-specific linear model, although this model did have the benefit of training on the validation dataset and requires manual image annotation. The demographics network does have significant appeal in not requiring ultrasound at all, which may be valuable in the clinical realm where patient demographics are always on file, but a current ultrasound

image may not be. The image and demographics model is also worth further exploration, although this exploration may only be justified if the size of the dataset grows significantly.

Table 12: Model summarization. *Comparison of model performance with mean absolute percent error.*

Model (Prior Work)	MAPE	Model (Thesis)	MAPE
All Location Statistical Linear Model	25.4%	Demographics Network	25.5%
Locations Combined Linear Model	26.8%	Demographics+Thickness Network	26.3%
All Locations Physics Linear Model	29.7%	Image Convolutional Network	31.6%
Locations Combined Physics Linear Model	26.8%	Image+Demograpihcs Convolutional Network	25.9%

With all the models clustered so tightly, with only a slight differences in mean absolute percent error (around 6%) it is important to consider what the mean absolute percent error levels mean. This is a challenging question, with implications on the relevance of the above models. A lack of repeated indentation trials at the same regions makes it difficult to ascertain whether a mean absolute percent error of 25% is significant or not. If a repeated indentation trial had a percent error of 25% on its own, it becomes reasonable to conclude that the models may have plateaued in performance and no improvements can be made as model error is only fitting to innate experimental error. On

the other hand, if the difference in stiffness between diseased tissue and healthy tissue is statistically small, as it is in some pathologies (Chokhandre et al., 2012), then the 25% percent error is unacceptably high and the models are not relevant for usage with their current data. Further experimentation and data collection may be needed to answer the above question.

This work could be extended in several possible ways, some of which are easier to implement than others. One of the simplest extensions would be to extend to convolutional neural network to other images during ultrasound indentation. Only the first image was used in an effort to explore whether tissue form was predictive of tissue indentation response. Incorporating data from the full indentation trial would provide more images for the model to train from and provide models with tissues in various levels of indentation. The current model was only exposed to minimally indented tissue and may not accurately predict surface compliance of indented tissue images, even though tissue surface compliance was considered to be a constant across indentation. Exploring non-linear mechanical behavior with a model as indentation occurs could provide a more detailed prediction of soft tissue behavior.

Another area of future work would be reevaluating the inputs to the convolutional neural network, by recreating the ultrasound images into a physical coordinate system. Ultrasound images were taken all at the same width (because the dimensions of the probe itself are constant), but the depth of tissue at each location was variable. The ultrasound system adaptively zooms in to show the entire tissue. As such, some tissue was presented in differing values of physical dimensions, which the model may not be aware of. Creating ultrasound images that are of variable height to represent the images in a

consistent physical coordinate system might improve model performance by representing tissue thickness within the image rather than just as a scaling factor fed into the model with demographics information. Adding image metadata may mitigate some of this, but reconstruction of the images would be the most representative way to feed the image into a machine learning workflow.

A possible interesting extension would be to use feature importance to find which parts of the ultrasound image led to predicting an image's compliance higher or lower. This would be similar to the SHAP plots in Chapter IV. Shapley or a similar library such as Alibi could be used (Klaise et al., 2021). This work may not necessarily improve convolutional neural network model performance but may provide insight into how the machine is making its predictions.

A final area of future work would be to expand the modeling work by exploring different machine learning models or finding methods to expand the datasets. This would be a large undertaking given initiating human subjects research is a time-consuming process. This experimentation could provide information on the repeatability of the tissue compliance measurement. Despite the effort expanding the data would involve, it may be the most worthwhile effort as the lack of data limits the usage of these models outside of research settings.

BIBLIOGRAPHY

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mane, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viegas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X., 2016. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems.

Abid, Abubakar, Abdalla, A., Abid, Ali, Khan, D., Alfozan, A., Zou, J., 2019. Gradio: Hassle-Free Sharing and Testing of ML Models in the Wild. https://doi.org/10.48550/arXiv.1906.02569

Albawi, S., Abed Mohammed, T., ALZAWI, S., 2017. Understanding of a Convolutional Neural Network. https://doi.org/10.1109/ICEngTechnol.2017.8308186

Barbieri, C., Cecatti, J.G., Souza, C.E., Marussi, E.F., Costa, J.V., 2008. Inter- and intra-observer variability in Sonographic measurements of the cross-sectional diameters and area of the umbilical cord and its vessels during pregnancy. Reprod. Health 5, 5. https://doi.org/10.1186/1742-4755-5-5

Benjamin, M., Kaiser, E., Milz, S., 2008. Structure-function relationships in tendons: a review. J. Anat. 212, 211–228. https://doi.org/10.1111/j.1469-7580.2008.00864.x

Bernal-Rusiel, J.L., Greve, D.N., Reuter, M., Fischl, B., Sabuncu, M.R., 2013. Statistical analysis of longitudinal neuroimage data with Linear Mixed Effects models. NeuroImage 66, 249–260. https://doi.org/10.1016/j.neuroimage.2012.10.065

Blackburn, H., Jacobs, D., Jr, 2014. Commentary: Origins and evolution of body mass index (BMI): continuing saga. Int. J. Epidemiol. 43, 665–669. https://doi.org/10.1093/ije/dyu061

Bonaccorso, G., 2020. Mastering Machine Learning Algorithms: Expert techniques for implementing popular machine learning algorithms, fine-tuning your models, and understanding how they work, 2nd Edition. Packt Publishing Ltd.

Brattain, L.J., Telfer, B.A., Dhyani, M., Grajo, J.R., Samir, A.E., 2018. Machine learning for medical ultrasound: status, methods, and future opportunities. Abdom. Radiol. 43, 786–799. https://doi.org/10.1007/s00261-018-1517-0

Choi, W.J., Russell, C.M., Tsai, C.M., Arzanpour, S., Robinovitch, S.N., 2015. Age-related changes in dynamic compressive properties of trochanteric soft tissues over the hip. J. Biomech. 48, 695–700. https://doi.org/10.1016/j.jbiomech.2014.12.026

Chokhandre, S., Halloran, J.P., van den Bogert, A.J., Erdemir, A., 2012. A Three-Dimensional Inverse Finite Element Analysis of the Heel Pad. J. Biomech. Eng. 134. https://doi.org/10.1115/1.4005692

Chollet, F., 2017. Xception: Deep Learning with Depthwise Separable Convolutions.

Coffin, C.T., 2014. Work-related musculoskeletal disorders in sonographers: a review of causes and types of injury and best practices for reducing injury risk. Rep. Med. Imaging 7, 15–26. https://doi.org/10.2147/RMI.S34724

Doherty, S., Landis, B., Owings, T.M., Erdemir, A., 2022. Template models for simulation of surface manipulation of musculoskeletal extremities. PLOS ONE 17, e0272051. https://doi.org/10.1371/journal.pone.0272051

Domingos, P., 2012. A few useful things to know about machine learning. Commun. ACM 55, 78–87. https://doi.org/10.1145/2347736.2347755

Erdemir, A., 2019. Reference Models for Multi-Layer Tissue Structures. Cleveland Clinic Cleveland United States.

Erdemir, A., Mulugeta, L., Ku, J.P., Drach, A., Horner, M., Morrison, T.M., Peng, G.C.Y., Vadigepalli, R., Lytton, W.W., Myers, J.G., 2020. Credible practice of modeling and simulation in healthcare: ten rules from a multidisciplinary perspective. J. Transl. Med. 18, 369. https://doi.org/10.1186/s12967-020-02540-4

Eskridge, S.L., Macera, C.A., Galarneau, M.R., Holbrook, T.L., Woodruff, S.I., MacGregor, A.J., Morton, D.J., Shaffer, R.A., 2012. Injuries from combat explosions in Iraq: Injury type, location, and severity. Injury 43, 1678–1682. https://doi.org/10.1016/j.injury.2012.05.027

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., Dean, J., 2019. A guide to deep learning in healthcare. Nat. Med. 25, 24–29. https://doi.org/10.1038/s41591-018-0316-z

Farage, M.A., Miller, K.W., Elsner, P., Maibach, H.I., 2013. Characteristics of the Aging Skin. Adv. Wound Care 2, 5–10. https://doi.org/10.1089/wound.2011.0356

Faraway, J.L., 2005. Extending the linear model with R: gerenalized linear, mexid effects an nonparametric regression models. Extending Linear Model R Gerenalized Linear Mexid Eff. Nonparametric Regres. Models 312–312.

Faustini, M.C., Neptune, R.R., Crawford, R.H., 2006. The quasi-static response of compliant prosthetic sockets for transtibial amputees using finite element methods. Med. Eng. Phys. 28, 114–121. https://doi.org/10.1016/j.medengphy.2005.04.019

Fithian, D.C., Kelly, M.A., Mow, V.C., 1990. Material properties and structure-function relationships in the menisci. Clin. Orthop. 19–31.

Gasparini, A., Abrams, K.R., Barrett, J.K., Major, R.W., Sweeting, M.J., Brunskill, N.J., Crowther, M.J., 2020. Mixed-effects models for health care longitudinal data with an informative visiting process: A Monte Carlo simulation study. Stat. Neerlandica 74, 5–23. https://doi.org/10.1111/stan.12188

Gilbertson, M.W., Anthony, B.W., 2013. An ergonomic, instrumented ultrasound probe for 6-axis force/torque measurement. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. IEEE Eng. Med. Biol. Soc. Annu. Int. Conf. 2013, 140–143. https://doi.org/10.1109/EMBC.2013.6609457

Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep Learning. MIT Press.

Hadda, V., Kumar, R., Dhungana, A., Khan, M., Madan, K., Khilnani, G., 2017. Inter- and intra-observer variability of ultrasonographic arm muscle thickness measurement by critical care physicians. J. Postgrad. Med. 63, 157–161. https://doi.org/10.4103/0022-3859.201412

Hara, K., Kataoka, H., Satoh, Y., 2017. Learning Spatio-Temporal Features with 3D Residual Networks for Action Recognition. https://doi.org/10.48550/arXiv.1708.07632

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep Residual Learning for Image Recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Presented at the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Las Vegas, NV, USA, pp. 770–778. https://doi.org/10.1109/CVPR.2016.90

He, K., Zhang, X., Ren, S., Sun, J., 2015. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification, in: 2015 IEEE International Conference on Computer Vision (ICCV). Presented at the 2015 IEEE International Conference on Computer Vision (ICCV), IEEE, Santiago, Chile, pp. 1026–1034. https://doi.org/10.1109/ICCV.2015.123

Heo, S.-J., Kim, Y., Yun, S., Lim, S.-S., Kim, J., Nam, C.-M., Park, E.-C., Jung, I., Yoon, J.-H., 2019. Deep Learning Algorithms with Demographic Information Help to Detect Tuberculosis in

Chest Radiographs in Annual Workers' Health Examination Data. Int. J. Environ. Res. Public. Health 16, 250. https://doi.org/10.3390/ijerph16020250
James, A.L., Palmer, L.J., Kicic, E., Maxwell, P.S., Lagan, S.E., Ryan, G.F., Musk, A.W., 2005. Decline in lung function in the Busselton Health Study: the effects of asthma and cigarette smoking. Am. J. Respir. Crit. Care Med. 171, 109–114. https://doi.org/10.1164/rccm.200402-230OC
Klaise, J., Looveren, A.V., Vacanti, G., Coca, A., 2021. Alibi Explain: Algorithms for Explaining Machine Learning Models. J. Mach. Learn. Res. 22, 1–7.
Knothe Tate, M.L., Gunning, P.W., Sansalone, V., 2016. Emergence of form from function—Mechanical engineering approaches to probe the role of stem cell mechanoadaptation in sealing cell fate. Bioarchitecture 6, 85–103. https://doi.org/10.1080/19490992.2016.1229729
Krizhevsky, A., Sutskever, I., Hinton, G.E., 2017. ImageNet classification with deep convolutional neural networks. Commun. ACM 60, 84–90. https://doi.org/10.1145/3065386
Lee, H.J., Yoon, D.Y., Seo, Y.L., Kim, J.H., Baek, S., Lim, K.J., Cho, Y.K., Yun, E.J., 2018. Intraobserver and Interobserver Variability in Ultrasound Measurements of Thyroid Nodules. J. Ultrasound Med. Off. J. Am. Inst. Ultrasound Med. 37, 173–178. https://doi.org/10.1002/jum.14316
Lento, P.H., Primack, S., 2007. Advances and utility of diagnostic ultrasound in musculoskeletal medicine. Curr. Rev. Musculoskelet. Med. 1, 24–31. https://doi.org/10.1007/s12178-007-9002-3
Li, X., Liu, Z., Cui, S., Luo, C., Li, C., Zhuang, Z., 2019. Predicting the effective mechanical property of heterogeneous materials by image based modeling and deep learning. Comput. Methods Appl. Mech. Eng. 347, 735–753. https://doi.org/10.1016/j.cma.2019.01.005
Lin, M., Chen, Q., Yan, S., 2014. Network In Network.
Liu, S., Wang, Y., Yang, X., Lei, B., Liu, L., Li, S.X., Ni, D., Wang, T., 2019. Deep Learning in Medical Ultrasound Analysis: A Review. Engineering 5, 261–275. https://doi.org/10.1016/j.eng.2018.11.020
Lundberg, S.M., Lee, S.-I., 2017. A Unified Approach to Interpreting Model Predictions, in: Advances in Neural Information Processing Systems. Curran Associates, Inc.
McInnes, E., Jammali-Blasi, A., Bell-Syer, S.E., Dumville, J.C., Middleton, V., Cullum, N., 2015. Support surfaces for pressure ulcer prevention. Cochrane Database Syst. Rev. https://doi.org/10.1002/14651858.CD001735.pub5
Mihai, L.A., Chin, L., Janmey, P.A., Goriely, A., 2015. A comparison of hyperelastic constitutive models applicable to brain and fat tissues. J. R. Soc. Interface 12, 20150486. https://doi.org/10.1098/rsif.2015.0486
Moerman, K.M., Vijven, M. van, Solis, L.R., Haaften, E.E. van, Loenen, A.C.Y., Mushahwar, V.K., Oomens, C.W.J., 2017. On the importance of 3D, geometrically accurate, and subject-specific finite element analysis for evaluation of in-vivo soft tissue loads. Comput. Methods Biomech. Biomed. Engin. 20, 483–491. https://doi.org/10.1080/10255842.2016.1250259
Neumann, E.E., Owings, T.M., Erdemir, A., 2019. Regional variations of in vivo surface stiffness of soft tissue layers of musculoskeletal extremities. J. Biomech. 95, 109307. https://doi.org/10.1016/j.jbiomech.2019.08.001
Neumann, E.E., Owings, T.M., Schimmoeller, T., Nagle, T.F., Colbrunn, R.W., Landis, B., Jelovsek, J.E., Wong, M., Ku, J.P., Erdemir, A., 2018. Reference data on thickness and mechanics of tissue layers and anthropometry of musculoskeletal extremities. Sci. Data 5, 180193. https://doi.org/10.1038/sdata.2018.193

Regueira, Y., Fargo, J., Tiller, D., Brown, K., Clements, C., Beacham, B., Brignone, E., Sommers, M., 2019. Comparison of Skin Biomechanics and Skin Color in Puerto Rican and Non-Puerto Rican Women. P. R. Health Sci. J. 38, 170–175.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation, in: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, Lecture Notes in Computer Science. Springer International Publishing, Cham, pp. 234–241. https://doi.org/10.1007/978-3-319-24574-4_28

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L., 2015. ImageNet Large Scale Visual Recognition Challenge. Int. J. Comput. Vis. 115, 211–252. https://doi.org/10.1007/s11263-015-0816-y

Satava, R.M., 1993. Virtual reality surgical simulator. Surg. Endosc. 7, 203–205. https://doi.org/10.1007/BF00594110

Schimmoeller, T., Cho, K.-H., Colbrunn, R., Nagle, T., Neumann, E., 2018. Instrumentation of Surgical Tools To Measure Load and Position During Incision, Tissue Retraction, and Suturing. Conf. Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. IEEE Eng. Med. Biol. Soc. Annu. Conf. 2018, 933–936. https://doi.org/10.1109/EMBC.2018.8512332

Schimmoeller, T., Colbrunn, R., Nagle, T., Lobosky, M., Neumann, E.E., Owings, T.M., Landis, B., Jelovsek, J.E., Erdemir, A., 2019. Instrumentation of off-the-shelf ultrasound system for measurement of probe forces during freehand imaging. J. Biomech. 83, 117–124. https://doi.org/10.1016/j.jbiomech.2018.11.032

Schmidhuber, J., 2015. Deep Learning in Neural Networks: An Overview. Neural Netw. 61, 85–117. https://doi.org/10.1016/j.neunet.2014.09.003

Shickel, B., Tighe, P.J., Bihorac, A., Rashidi, P., 2018. Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. IEEE J. Biomed. Health Inform. 22, 1589–1604. https://doi.org/10.1109/JBHI.2017.2767063

Sigrist, R.M.S., Liau, J., Kaffas, A.E., Chammas, M.C., Willmann, J.K., 2017. Ultrasound Elastography: Review of Techniques and Clinical Applications. Theranostics 7, 1303–1329. https://doi.org/10.7150/thno.18650

Simonyan, K., Zisserman, A., 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. https://doi.org/10.48550/arXiv.1409.1556

Sippel, S., Muruganandan, K., Levine, A., Shah, S., 2011. Review article: Use of ultrasound in the developing world. Int. J. Emerg. Med. 4, 72. https://doi.org/10.1186/1865-1380-4-72

Speck, T., Burgert, I., 2011. Plant Stems: Functional Design and Mechanics. Annu. Rev. Mater. Res. 41, 169–193. https://doi.org/10.1146/annurev-matsci-062910-100425

Tang, H., Buehler, M.J., Moran, B., 2009. A Constitutive Model of Soft Tissue: From Nanoscale Collagen to Tissue Continuum. Ann. Biomed. Eng. 37, 1117–1130. https://doi.org/10.1007/s10439-009-9679-0

Topol, E., 2019. Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again. Basic Books.

U.S. Census Bureau QuickFacts: United States [WWW Document], n.d. URL https://www.census.gov/quickfacts/fact/table/US/PST045221 (accessed 9.28.22).

Volovici, V., Syn, N.L., Ercole, A., Zhao, J.J., Liu, N., 2022. Steps to avoid overuse and misuse of machine learning in clinical research. Nat. Med. 1–4. https://doi.org/10.1038/s41591-022-01961-6

Wang, J.-M., Luo, X.-N., Li, Y., Dai, X.-Q., You, F., 2016. The Application of the Volumetric Subdivision Scheme in the Simulation of Elastic Human Body Deformation and Garment Pressure: Text. Res. J. https://doi.org/10.1177/0040517505054174

Weisstein, E.W., n.d. Convolution [WWW Document]. URL https://mathworld.wolfram.com/ (accessed 10.13.22).

Wesley, N.O., Maibach, H.I., 2003. Racial (ethnic) differences in skin properties: the objective data. Am. J. Clin. Dermatol. 4, 843–860. https://doi.org/10.2165/00128071-200304120-00004

Winder, M., Owczarek, A.J., Chudek, J., Pilch-Kowalczyk, J., Baron, J., 2021. Are We Overdoing It? Changes in Diagnostic Imaging Workload during the Years 2010–2020 including the Impact of the SARS-CoV-2 Pandemic. Healthcare 9, 1557. https://doi.org/10.3390/healthcare9111557

APPENDIX A

GITHUB REPOSITORY STRUCTURE

Link: https://github.com/sbdoherty/UltrasoundML

doc:

"doherty_thesis.docx" – Thesis document

dat:

001_MasterList_indentation_orig.csv – Data for training all models

src:

"MixedEffectModel - MS_Thesis.R" – Chapter III models

demographics_DNN.py – Chapter IV, demographics network

image_based_DNN.py – Chapter IV, demographics+thickness network

image_CNN.py – image only convolutional neural network

image_and_demographics_CNN.py – image, demographics, and metadata network