LEVERAGING MACHINE AND DEEP LEARNING TO DEVELOP OPTIMIZED AND NOVEL MODELS FOR PATIENTS UNDERGOING ABDOMINAL WALL RECONSTRUCTION SURGERY

by

Keith Joseph Murphy

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Health Services Research

Charlotte

2023

Approved by:

_____
Dr. Rajib Paul

_____
Dr. B. Todd Heniford

_____
Dr. Shi Chen

_____
Dr. Hamed Tabkhi

_____
Dr. Wlodek Zadrozny

ABSTRACT

KEITH MURPHY. Leveraging Machine and Deep Learning to Develop Optimized and Novel Models for Patients Undergoing Abdominal Wall Reconstruction Surgery.
Under the direction of DR. RAJIB PAUL

This research investigates the intersection of applied statistics, machine learning, and healthcare to leverage the quantitative approaches as a means of improving patient quality of life after surgery. Relevant patient data prior to surgery remains relatively limited. Therefore, finding the appropriate means to maximize the limited data available to optimize patient outcomes postoperatively is imperative. This study utilizes multiple patient cohorts, drawn from a tertiary care hernia referral center in the southeastern United States, who underwent abdominal wall reconstruction surgery.

Chapter 2 focuses on developing a multivariable model using unique preoperative patient features to model the relationship between these variables and the outcome of interest: patient quality of life six months following abdominal wall reconstruction surgery. Using patient cohort data from the years 2005-2017, this study successfully built and internally validated a multivariable model using 20 unique preclinical variables. These findings provide further evidence to determine what preoperative variables reliably predict patient quality of life after surgery. Ultimately, this assists clinicians in preoperative assessments to optimize early patient interventions and treatment plans.

Chapter 3 shifts to investigating alternate data sources in predicting patient outcomes, i.e., patient imaging data in the form of computed tomography scans. Since these data are scarce prior to surgery, this research focuses on assessing multiple qualitative methodologies to align and improve image quality for predictive modeling. From the various methodologies analyzed, image averaging after cropping and alignment showed the most promising means of optimizing

preoperative patient images for patient quality of life classification. Using these techniques aids researchers in maximizing the limited image data available for building accurate classification models in surgery.

Finally, Chapter 4 explores existing predictive models in abdominal wall reconstruction surgery to determine external generalizability on other patient cohorts. This research draws on the methods and techniques explored in the third chapter to optimize patient images to successfully train and validate these models. Although not initially successful in demonstrating model external validity on outside patients, investigation in pooled validation techniques suggests successful and generalizable models are possible with further investigation into matching internal and external patient cohorts. In conclusion, this research explores the possible applications of statistical and machine learning methods in surgery and provides means of implementing these techniques successfully in the clinical context.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to Dr. Rajib Paul who served as the chair and primary advisor of my dissertation. His contributions were immeasurable, and without his help, this dissertation would not have been possible. My schedule required me to work full-time while completing this dissertation and Dr. Paul brought assurance and flexibility in times where stress for me both academically and professionally looked insurmountable. Dr. Paul never hesitated to generously contribute his time and effort through each step of this dissertation. His contributions and insights into each chapter provided invaluable mentorship in teaching not only the statistical and research techniques required to conduct the research, but more importantly, the significance and understanding behind why we pursued such methods. He always went out of his way to provide opportunities for academic growth through abstract submissions, external projects, and academic competitions to apply the knowledge gained during my dissertation. I hope that he continues to inspire and instruct other students like myself moving forward in his academic career.

I also want to acknowledge Drs. Heniford, Chen, Tabkhi, and Zadrozny who supported and assisted me through this dissertation process as part of my committee. Each individual provided critical insight and knowledge in their respective fields of expertise which has only served to deepen my understanding of applied statistics and machine learning within the context of surgery. Of special note, I would like to thank Dr. Heniford and his team at Atrium Health for providing the data used throughout this dissertation. I hope that this dissertation can contribute in furthering the research in improving patient outcomes after surgery.

DEDICATION

This dissertation is dedicated to my mother and father, who taught me the value of education and always letting my conscience be my guide.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

MIS An acronym for Minimally Invasive Surgery

QOL An acronym for Quality of Life

CT An acronym for Computed Tomography

DLM An acronym for Deep Learning Model

CNN An acronym for Convolutional Neural Network

ROC An acronym for Receiver Operating Characteristic

AWR An acronym for Abdominal Wall Reconstruction

CCS An acronym for Carolinas Comfort Scale

SF-36 An acronym for the Short-Form 36 survey

PPP An acronym for Persistent Postoperative Pain

MICE An acronym for Multiple Imputation by Chained Equations

SAENET An acronym for Stacked Adaptive ElasticNet

IQR An acronym for Interquartile Range

BMI An acronym for Body Mass Index

MRI An acronym for Magnetic Resonance Imaging

ILSVRC An acronym for the ImageNet Large Scale Visual Recognition Challenge

SGD An acronym for Stochastic Gradient Descent

DLR An acronym for Deep Learning Reconstruction

AI An acronym for Artificial Intelligence

ANN An acronym for Artificial Neural Network

AUROC An acronym for Area Under the Receiving Operating Characteristic

CST An acronym for Component Separation Technique

SSI An acronym for Surgical Site Infection

LOOCV An acronym for Leave-One-Out Cross-Validation

CHAPTER 1: INTRODUCTION

Minimally Invasive Surgery and Hernia Repairs

The 1980s marked a significant event in the advancement of surgical techniques: the introduction of minimally invasive surgery (MIS). This new suite of techniques provided surgeons with a means of operating on patients with significantly less risk of damage to the patient's body, as compared to open surgery (Schlich & Tang, 2017). As a result, MIS techniques have been adopted and incorporated into multiple surgical domains including spinal, abdominal, urological, and hepato-pancreato-biliary surgeries. Surgical repairs for abdominal surgery represent some of the most commonly performed procedures in MIS, with approximately 350,000 ventral and incisional hernia repairs occurring annually in the United States (Wechter et. al., 2005).

From an economic perspective, the estimated healthcare costs for such procedures totaled in excess of $3.2 billion in 2006 and continue to rise each year (Poulose et. al., 2012). Proper management following procedures is imperative in reducing adverse postoperative outcomes and hernia recurrence. For each percentage increase in average population hernia recurrence, Poulose et. al., (2012) estimated additional healthcare costs at $32 million dollars. With significant improvements made to reducing mortality rates across surgical domains in the past 70 years, the focus has now shifted to minimizing postoperative morbidity rates and optimizing patient quality of life (QOL) after surgery (Heniford et. al., 2008).

Quality of Life in Surgery

As previously discussed, health-related QOL has become an increasingly utilized measure to evaluate the outcome of surgical care. This metric encompasses a wide range of concepts including a patient's physical health, psychological state, social relationships, and level

of autonomy (Urbach, 2005). Not all of these aspects are applicable to surgery and surgery research. More specifically, those appropriate in describing QOL are broadly defined as patient-based outcome measures. In essence, the goal of surgical care is inextricably tied to patient QOL. As the primary purpose of surgical care is to improve a patient's health, the interventional effect surgery has on patient QOL could be argued as the most significant metric in evaluating the efficacy and effectiveness of surgical care (Urbach, 2005). However, due to the complexity inherent in defining patient QOL, one cannot directly measure this phenomenon in a patient (Fitzpatrick et. al., 1998). The most common estimation for QOL involves utilizing a derived measurement calculated from a questionnaire or survey. For example, a QOL measure might seek to quantify chronic pain in patients who suffer from gastroesophageal reflux disease (GERD) (Velanovich, 1998).

In MIS, rates for chronic pain following hernia repair have been reported as high as 39%, demonstrating a QOL measure that necessitates further investigation to optimize preoperative interventions and ultimately reduce patient pain following surgery (Luijendijk et. al., 2000). Traditionally, QOL instruments in medicine were primarily designed to measure multiple aspects of health, the most notable example in health research literature being the short-form 36 (SF-36) survey (Ware & Sherbourne, 1992). This 36-item questionnaire includes a variety of questions designed to calculate multiple measurements of patient QOL including physical, mental and emotional health.

Predictive Models in Surgery

Due to this shift towards quantifying patient QOL after surgery, many surgeons have sought to identify what patient features, if any, could be associated with this outcome. This can include, but is not limited to: patient demographics, co-morbidities, lab measurements,

diagnoses, interoperative factors, and immediate postoperative factors (Lee et. al., 2016). This development catalyzed researchers into developing predictive algorithms and formulae capable of modeling these relationships between these patient features and characteristics and the outcome in question. This surge is exemplified by the number of research publications in medicine involving predictive modeling increasing exponentially over the past two decades (Bendifallah et. al., 2015).

Predictive modeling provides the means of stratifying patients based on risk for adverse outcomes (Osorio et. al., 2016). Among predictive modeling techniques, perhaps the most ubiquitously seen in medical literature is logistic regression (Hosmer & Lemeshow, 2013). This method is commonly used to model the probabilistic relationship between independent features on an adverse health outcome (Chen & Yun-Fang, 2013). This research has resulted in improved standardization processes across surgical domains with the intent of prioritizing both patient safety and postoperative outcomes. Ultimately, the ability to provide accurate predictions of patients' postoperative outcomes has significant implications for tailored preoperative counseling, and more effective interventional and treatment plans (Bekelis et. al., 2014). These predictive models can facilitate more objective means of determining patients at high risk for adverse outcomes, and thus provide additional tools for more informed clinical decision-making with the intent of improving patient health and quality of life.

Regarding MIS and hernia repair, numerous efforts have been made to model associations of singular risk factors with various postoperative surgical outcomes through multivariate analyses and logistic regression-based approaches (Montes et. al., 2015, Schug & Bruce, 2016). To date, no studies have aimed to predict patient postoperative QOL following hernia surgery utilizing purely preoperative patient features. Chapter Two will outline the current

state of disease-specific QOL measures in hernia surgery as well as the research in existing predictive models in MIS. Furthermore, the chapter will investigate generating a novel predictive algorithm built utilizing patient preoperative factors to model the association between these features and patient postoperative QOL.

Deep Learning: A Brief Primer

With the widespread success and numerous publications using predictive modeling in medicine, researchers began to investigate what improvements could be made to existing modeling techniques to provide more accurate, valid, and generalizable predictions based on the wealth of patient data available. In the last decade, a newer algorithm gained notoriety in the clinical research community: deep learning. This methodology was of significant interest to medical researchers, as it provided a means of modeling complex data inputs i.e., text and image data which were previously unable to be analyzed adequately by more traditional statistical and machine learning modeling techniques (Wang et. al., 2018). In MIS, the majority of preoperative imaging consists of computed tomography (CT) images which enable trained clinicians to investigate internal anatomical structures and possible issues within a patient (Zhu et. al., 2022). These scans have been traditionally employed to determine clinical diagnosis and treatment courses for patient hernias.

Before delving into the deep learning applications in MIS, we must first define deep learning models (DLM) and their general purpose and applications. The DLM is a subset of machine learning originally inspired to mimic the function of a neuron. The model takes an input that undergoes multiple transformations to provide a data representation tailored to solving a predetermined problem (Rivas-Blanco, et. al., 2021). Some of the most well-known and documented DLMs are those used in image processing applications which include AlexNet,

VGGNet, and ResNet. These DLMs are normally referred to as Convolutional Neural Networks (CNNs) and follow a general architecture of a series of convolutional and pooling layers as well as a fully connected layer (Shrestha & Mahmood, 2019). The convolution layers are where the model learns the localized patterns among an image's pixels to extract high-level feature information. These layers can capture such information as an image's edges or colors, to higher level representations within an image, such as the detected object of interest in an image. Next, the pooling layers decrease the dimensionality of the features generated from convolutional layers to summarize the information and consequently make them more robust to any possible feature variations that may occur in the image (Chollet, 2017). Finally, the fully connected layer provides the actual classification for our image input which is accomplished by learning non-linear feature combinations provided through the pooling layer.

Designing an ideal CNN architecture for research classification is not an easy task, as the deep learning architect must decide on the optimal number of layers in the DLM. This choice comes with multiple trade-offs including computational speed, good model classification performance, and adequate generalization ability on new input data (Jia et. al., 2009). In practice, having a "shallow" network with fewer convolutional layers will generally perform predictions much faster and with better generalizability but with less accuracy. In contrast, a deeper network which has the ability to model greater numbers of parameters will on average have much better prediction scores compared to shallower networks on the training set; however, this requires much greater processing time at the expense of generalizability and with the added risk of overfitting on the training data (Xiao et. al., 2020). Another key element to developing an accurate predictive model includes having large amounts of labeled data for model training so that the model can "learn" the inherent variability that could be present among the true

population of images. Obtaining sufficiently large quantities of images for adequate training can represent a daunting task, however, and significant research has been invested into determining the sufficient number of images to adequately train a deep learning model for classification (Simonyan & Zisserman, 2014).

Limitations of Deep Learning in Medicine

As previously mentioned, obtaining sufficient quantities of annotated data in specific research domains like MIS is a difficult and costly process. Building deep learning models that can adequately classify input data requires a significant amount of training images, with some researchers hypothesizing at least 10,000 unique images per class as a necessity for robust model development (Akkus et. al., 2017). Furthermore, many research questions in surgery involve rare disease states, procedures, and outcomes, which can result in very skewed datasets where positive or experimental cases represent as little as 5% of all sampled individuals. These heavily skewed datasets pose a difficult problem for deep learning classification, as it is unlikely to reach the requisite number of images to capture the full variation possible in positive cases. Utilizing clinical data also comes with other complications, including legal agreements over utilizing protected patient information, cost of requiring data, and, as previously mentioned, significant time and effort in annotating and labeling training classes for images (Shorten et. al., 2019).

Due to these limitations, clinical researchers have sought methods to combat these issues inherent in applying these classification models in surgery. Consequently, two major methods have arisen to address data limitations which include quantity- and quality-driven approaches to data augmentation. Broadly speaking, data augmentation encompasses a host of techniques concerned with optimizing an existing, albeit limited, image dataset to more accurately represent the true variance present in the larger image population. Quantity-driven approaches to

augmentation use transformational techniques to artificially inflate the existing image sample size to capture the variance possible between images in a given dataset (Lo et. al., 2021). These include, but are not limited to: rotating an image, flipping an image on one or multiple axes, or image magnification or shrinking. For qualitative improvements, these generally include modifying pixel intensities to particular thresholds to reduce background noise which could result in image misclassification (Balaji & Sumathi, 2014).

In MIS and hernia repair, similar to other surgical domains, many of the positive postoperative adverse outcomes are relatively limited. For example, rates of surgical site infection following hernia surgery are reportedly <10% which, as previously mentioned, provides a difficult outcome to classify using deep learning primarily due to the dataset's class imbalances (Elhage et. al., 2021). Although there has been considerable discussion on these limitations in utilizing surgical data for image classification, there is limited available research on methods to optimize these images prior to model input and processing for analysis (Chunwei et. al., 2020).

Chapter Three of this paper is concerned with optimizing image quality for building more accurate and discriminative predictive models. This chapter will investigate utilizing multiple quality-driven approaches to data augmentation to determine if these techniques can provide any appreciable difference in model accuracy as compared to utilizing the original, unmodified dataset.

Validating Deep Learning Models in Medicine

One of the earliest examples of deep learning applications utilizing medical imaging to classify outcomes was published in *Nature* by Wang et. al. in 2016. This paper sought to classify

breast macrocalcifications as either malignant or benign based on preoperative screening images of patient mammograms. The DLM outperformed all other classification algorithms with higher predictive accuracy and receiver operating characteristic (ROC) values. Deep learning publications in surgery continue to grow, with greater numbers of papers published utilizing deep learning for classification in medicine.

In regards to MIS, Elhage et. al., (2021) published a study concerning the implementation of a DLM successfully on a patient cohort who underwent abdominal wall reconstruction (AWR) surgery to classify procedural difficulty and patient postoperative outcomes. Model ROC values for each outcome demonstrated strong discriminative ability (reported ROC values for risk of surgical complexity and wound infection as 0.74 and 0.89, respectively) and became one of the first studies to predict surgical complexity and postoperative outcomes using solely preoperative imaging data.

Although publication rates for predictive models and deep learning-based classification models are high, there is a dearth of literature on sufficiently validating these models with the intent of clinical implementation. Of the 85,000 available predictive model publications on Pubmed, less than 5% involved any form of external validation to determine model generalizability on cohorts outside of the initial training sample (Ramspek et. al., 2021). Model performance has historically been demonstrated as worse in external samples due to the potential of overfitting on the initial training sample, and therefore, those models without sufficient validation cannot be reasonably recommended for clinical implementation. Chapter Four will describe this gap in the literature on insufficient validations involving deep learning prediction. In addition, this chapter will aim to validate the DLM model generated by Elhage et. al. (2021) using an external patient cohort to further validate model accuracy, discriminative ability

(through receiver operating characteristic [ROC] value) and, ultimately, model generalizability on cohorts outside of the patient sample originally leveraged to generate the model.

CHAPTER 2: PREDICTING POSTOPERATIVE PATIENT QUALITY OF LIFE AFTER HERNIA SURGERY USING PREOPERATIVE FACTORS

2.1 Introduction

With significant improvements made in reducing morbidity and mortality rates after hernia surgery and repair, clinical focus has shifted to improvements in the patient's perceived quality of life (QOL) following hernia surgery (Heniford et. al., 2008). QOL assessments provide a means of assessing the repercussions following the medical conditions and treatments from the patient's perspective (Urbach, 2005). The Carolinas Comfort Scale (CCS) was originally generated in 2008 as a means of more accurately determining a patient's QOL following abdominal wall reconstruction (AWR) surgery (Heniford et. al., 2008). The score is calculated on a six-point scale based on answers from a 24-item questionnaire. Traditionally, the Short-Form 36 (SF-36) survey was utilized to assess patients' QOL after hernia surgery (Guyatt et. al., 1993). However, previous research notes that the SF-36 survey was unable to accurately assess patients' QOL compared to other clinically developed disease-specific questionnaires (Velanovich, 1998). Heniford and colleagues invented the CCS score as a reliable and accurate predictor of patient QOL after ventral and inguinal hernia surgery with a high Cronbach's α (0.97), test-retest validity, concurrent validity, and spearman's correlation coefficient demonstrating individual question significance within the survey. Furthermore, patient satisfaction in taking the CCS survey was significantly greater than in the comparative SF-36 survey (Heniford et. al., 2008).

In the past decade, the CCS score has been extensively validated both nationally and internationally across multiple hernia patient cohorts (Parseliunas et. al., 2022). Studies demonstrated high internal consistency with Cronbach $\alpha > 0.90$, construct validity, and patient survey satisfaction similar to those documented in the initial study. Due to the extensive validity

and widespread usage of the CCS score following hernia surgery, the instrument represented an ideal measure for determining patient QOL following hernia surgery.

Optimizing patient-centered surgical care involves many potential tradeoffs including benefits, risks, and costs when determining a patient's treatment plan (Wennberg, 2004). Ideally, all necessary information would be available to both the clinician and the patient before surgery to optimize patient care plans. However, current medical practices do not consider this comprehensive data-driven approach as necessary (Rudmik et. al., 2015). Therefore, identifying what preoperative factors are most significantly related to the patient's postoperative QOL could represent a significant decision-support tool in tailoring a patient's treatment plan and immediate postoperative interventions to maximize patients' QOL after surgery.

In MIS, hernia repairs are often difficult to effectively manage postoperatively (Borad & Merchant, 2017). Van Ramhorst and colleagues remarked that many patients go on to develop incisional hernias post-operatively which have significant negative impacts on both patients' QOL and patient perceived body image (van Ramhorst et. al., 2012). Among the 73 patients with incisional hernias, the authors noted a statistically significant reduction in SF-36 survey categories: "physical functioning" and "physical role" compared to those patients without incisional hernia at the time of primary surgery. Other possible negative clinical outcomes include persistent postoperative pain which can affect up to one-third of patients following inguinal hernia surgery. Furthermore, Vad et. al. noted that anywhere from 2-30% of patients will report persistent postoperative pain (PPP) within 6 months following hernia repair surgery (Vad. Et. al., 2011). The researchers noted pharmacological and clinical management of PPP to be extremely difficult with inconclusive results on interventional efficacy from other reported studies (Kehlet et. al., 2008).

Due to the aforementioned suite of postoperative negative outcomes associated with poor management of hernia care, the developing a predictive model that can stratify patients on postoperative QOL based on preoperative features is essential in optimizing care treatment after hernia surgery. Previous studies have sought to associate singular risk factors with various postoperative surgery outcomes through multivariable analyses and regression-based approaches. However, no studies have aimed to predict the patient postoperative QOL after hernia surgery solely using patient preoperative characteristics (Liang et. al., 2015, Schug & Bruce, 2016, Montes et. al., 2015).

Our aim is to identify individual and combinations of patient preoperative features either statistically or clinically associated with patient postoperative QOL, represented by the individual CCS score, six months following abdominal wall reconstruction (AWR) surgery. Clinically and statistically significant features will be utilized to build a multivariate model outlining the aforementioned association. In addition, we aim to develop a novel predictive modeling algorithm using this multivariate model to identify high-risk patients who may have significantly reduced QOL postoperatively.

2.2 Methods

Study Data Overview

This study incorporated data on $n = 250$ patients who underwent AWR surgery during the period of 2005-2017 at a tertiary care hernia referral center in the southeastern United States. Data included multiple features on patients' pre-, intra-, and post-operative information regarding surgery totaling $k = 774$ variables. Since this study is concerned with identifying what preoperative features are most significantly associated with the postoperative QOL outcome, any features that did not meet these criteria (i.e., those recorded intra- or post-operatively) were

removed in the data processing step. All statistical analyses were performed in Python v3.7.1 and R version 4.0 (Van Rossum & Drake, 2009, RStudio Team, 2019). All packages and dependencies utilized for generating the code are outlined in Appendix A.

Inclusion and Exclusion Criteria:

Patients analyzed in this study included those who underwent a ventral or incisional hernia repair procedure, using either open or laparoscopic approaches. Patients with missing outcome data, i.e., no CCS score at time of survey follow-up (0.6%) at six months were excluded from the analysis. In addition, this study included males or females 18 years or older as well as those with a literacy level capable of understanding registry patient questionnaires.

Outcome

The outcome variable, the CCS Score (ranging from 0-6), was calculated for each patient within the sample after a six-month follow-up survey following AWR surgery. CCS Scores were dichotomized to patients with a score ≥2, or <2. Previous studies by Cox and colleagues (2016) had outlined this variable binarization for analysis with CCS score of 0 (no symptoms) or 1 (minimal and not bothersome) as “asymptomatic patients” while those with CCS scores of 2 (signified as mild symptoms, but bothersome) or higher as “symptomatic patients.

Data Processing

Preoperative patient variables included multiple patient factors including demographics, preoperative lab values, co-morbidities, previous surgeries and complications, and preoperative QOL (quantified by CCS score). After eliminating intraoperative and postoperative variables, those remaining were assessed for missing values. Any variables with >50% of values representing missing or null values were removed from the analysis. Furthermore, any features with less than 5% positive cases were also removed during data processing. Feature correlations

were calculated using Pearson's correlation coefficient and any variable sets with an absolute correlation >0.7 (denoting "strong" correlation) were assessed by a surgeon expert and either removed or kept for further analysis (Assunção, et. al., 2017). Categorical variables were encoded using one-hot encoding methods to generate binary variables for each category. Due to the large date range encompassing patients in the sample, procedure year was also included as a variable to control for any major changes in surgical technique or practice over the past decade.

Data Analysis

Data were summarized by medians with interquartile ranges (for continuous variables) and frequencies (for binary and categorical variables). Independent variables were stratified by the outcome of interest and bivariate analyses were performed using Wilcoxon rank-sum test Chi-square tests of comparison or non-parametric Fisher's exact tests were also performed for binary variables. All tests of comparison were reported with the associated significance P-values. Following bivariate analysis, missing values among variables were imputed using multiple imputations by chained equation (MICE) forest techniques (Mera-Garona et., al., 2013) which generated 50 imputed datasets for filling missing values. Variable selection was performed using stacked adaptive elastic net (SAENET), a penalized regression technique that regularizes model parameters by shrinking the regression coefficients (Du. et. al., 2020). To calculate the associated penalization factor for features, SAENET utilizes an alpha/lambda pairing combined with a fivefold cross validation. Alpha controls the relative balance between reducing feature importance to small values or reducing feature values to zero. Lambda is calculated as one standard error from the minimum error determined from the cross validation. Taken together, the chosen pair is determined based on the highest L1 penalty and incorporated into the equation. All remaining non-zero features are therefore incorporated into the model.

Traditionally, combining multiple imputation with variable selection approaches is difficult as applying variable selection algorithms to each imputed dataset through multiple imputation can lead to varying sets of selected predictors and feature importance. We incorporated the “Miselect” R package which combines stacked adaptive elastic Net with forced variable selection across multiple imputed data (Du et. al., 2020). The algorithm pools objective functions across these imputations, and subsequently performs optimization jointly over all imputed datasets rather than separately for each dataset.

All remaining imputed independent variables were then incorporated into a multivariable logistic regression model. Models and features were assessed using model goodness of fit through adjusted R-squared, expert surgical opinion, and variable statistical significance level <0.05. Model goodness-of-fit was also assessed by Hosmer-Lemeshow tests to determine the model was a good fit for the data with a reported P-value of >0.05 (Hosmer & Lemeshow, 2013).

In addition, a second, complete-case dataset was generated for comparison with the candidate multivariable model. This complete-case dataset utilized only the candidate independent features identified from the Miselect package. In this case, all patient rows with missing data were dropped from the analysis table and the remaining data were fitted to the outcome. Significant variables in the non-imputed model were compared to the imputed multivariable model variables to assess congruency in results.

Candidate final models were validated by constructing training and test samples (80:20 train/test split) from the initial 250-sample patient cohort. Finally, models were evaluated on test sample receiver operating characteristic (ROC) scores to determine model predictive ability on patient CCS score six months after surgery. Following test ROC calculation, the patient sample

was further validated by 5-fold stratified sampling techniques to additionally cross-validate ROC values across 5 test samples using the built multivariate model (Rois et. al., 2021). The full data preprocessing steps and model analytics are summarized in Appendix B.

Table 2.1: Median and Proportion values for all candidate model variables stratified by CCS Score

Variable	Total (%) (n=250)	CCS Score <2 (%) (n=128)	CCS Score ≥2 (%) (n=122)	p-value
Patient Age	61.0 [52.0, 69.0]	63.0 [52.0, 69.0]	60.0[52.0, 69.0]	*0.242*
Patient Gender				
. Male	106(42.4)	67(52.3)	39(32.0)	***0.001****
. Female	144(57.6)	61(47.7)	83(68.0)	
Patient Race				
. White or Caucasian	230 (92.0)	120(93.7)	110 (90.2)	*0.417*
. Non-White	20 (8.0)	8 (6.3)	12(9.8)	
Patient BMI	31.6 [28.3,35.5]	30.4 [27.6, 34.6]	33.1 [29.1, 37.1]	***0.009****
Patient Taking Steroids				
. No	233(94.0)	120(95.2)	113(92.6)	*0.550*
. Yes	15(6.0)	6(4.8)	9(7.4)	
Patient Taking Coumadin				
. No	233(94.3)	122(96.8)	111(91.7)	*0.102*
. Yes	14(5.7)	4(3.2)	10(8.3)	
Patient Taking PPI				
. No	154(61.8)	81(63.8)	73(59.9)	*0.610*
. Yes	95(38.2)	46(36.2)	49(40.1)	
Patient Taking Antiplatelet				
. No	189(76.2)	92(72.4)	97(80.2)	*0.201*
. Yes	59(23.8)	35(27.6)	24(19.8)	
--Comorbidities--				
Alcoholism				

. No	246(98.8)	125(97.7)	121(100.0)	*0.248*
. Yes	3(1.2)	3(2.3)	0(0.0)	
Tobacco Use				
. No	218(87.2)	118(92.2)	100(82.0)	***0.025****
. Yes	32(12.8)	10(7.8)	22(18.0)	
Heart Arrythmias				
. No	240(96.4)	124(96.9)	116(95.9)	*0.743*
. Yes	9(3.6)	4(3.1)	5(4.1)	
Asthma				
. No	223(89.6)	120(93.7)	103(85.1)	***0.043****
. Yes	26(10.4)	8(6.3)	18(14.9)	
Congestive Heart Failure				
. No	245(98.4)	127(99.0)	118(97.5)	*0.358*
. Yes	4(1.6)	1(1.0)	3(2.5)	
Cirrhosis or Liver Disease				
. No	241(96.8)	122(95.3)	119(98.3)	*0.283*
. Yes	8(3.2)	6(4.7)	2(1.7)	
Chronic Obstructive Pulmonary Disease				
. No	239(96.0)	128(100.0)	110(91.7)	***<0.001****
. Yes	10(4.0)	0(0.0)	10(8.3)	
Coronary Artery Disease				
. No	224(90.0)	114(89.1)	110(90.9)	*0.784*
. Yes	25(10.0)	14(10.9)	11(9.1)	
Cerebrovascular Accident (Stroke)				
. No	240(96.4)	121(92.0)	119(98.7)	*0.173*
. Yes	9(3.6)	7(8.0)	2(1.3)	
Diabetes				
. No	191(76.4)	105(82.0)	86(70.5)	***0.046****
. Yes	59(23.6)	23(18.0)	36(29.5)	
End Stage Renal Disease				
. No	242(97.6)	125(98.4)	117(96.6)	*0.437*
. Yes	6(2.4)	2(1.6)	4(3.3)	

Gastroesophageal Reflux Disease (GERD)				
. No	173(77.2)	93(78.8)	80(75.5)	*0.663*
. Yes	51(22.8)	25(21.2)	26(24.5)	
History of Cancer				
. No	181(72.7)	87(68.0)	94(77.7)	*0.115*
. Yes	68(27.3	41(32.0)	27(22.3)	
Hypercholesterolemia				
. No	191(76.7)	95(74.2)	96(79.3)	*0.421*
. Yes	58(23.3)	33(25.8)	25(20.7)	
Hyperlipidemia				
. No	205(82.3)	103(80.5)	102(84.3)	*0.532*
. Yes	44(17.7)	25(19.5)	19(15.7)	
Hypertension				
. No	115(46.2)	53(41.1)	62(51.2)	*0.153*
. Yes	134(53.8)	75(58.9)	59(48.8)	
Hypotension				
. No	247(99.2)	126(98.4)	121(100.0)	*0.500*
. Yes	2(0.8)	2(1.6)	0(0.0)	
Hypothyroidism				
. No	221(88.8)	113(88.3)	108(89.3)	*0.966*
. Yes	28(11.2)	15(11.7)	13(10.7)	
Pre-operative Anemia				
. No	238(95.6)	120(93.7)	118(97.5)	*0.218*
. Yes	11(4.4)	8(6.3)	3(2.5)	
Previous Intra-abdominal Surgery or Trauma				
. No	7(2.8)	4(3.1)	3(2.5)	*0.940*
. Yes	242(97.2)	124(96.9)	118(97.5)	
Pulmonary Hypertension				
. No	247(99.2)	127(99.0)	120(99.0)	*1.000*
. Yes	2(0.8)	1(1.0)	1(1.0)	
Peripheral Vascular Disease				

. No	246(98.8)	127(99.0)	119(98.3)	*0.613*
. Yes	3(1.2)	1(1.0)	2(1.7)	
Sleep Apnea				
. No	216(86.7)	110(85.9)	106(87.6)	*0.713*
. Yes	33(13.3)	18(14.1)	15(12.4)	
Renal Insufficiency				
. No	245(98.8)	125(98.4)	120(99.0)	*1.000*
. Yes	3(1.2)	2(1.6)	1(1.0)	
--Lab Values--				
Sodium Levels	139.0[138.0,141.0]	140.0[138.0,141.0]	139.0[138.0,140.0]	*0.265*
Creatinine Levels	0.9 [0.7,1.1]	0.9 [0.7, 1.1]	0.9 [0.7, 1.1]	*0.738*
--Prev. Abdominal Surgery--				
Hysterectomy				
. No	178(71.2)	101(78.9)	77(62.8)	***0.009****
. Yes	72(28.8)	27(21.1)	45(37.2)	
Colectomy				
. No	177(70.8)	94(73.4)	83(68.0)	*0.423*
. Yes	73(29.2)	34(26.6)	39(32.0)	
Cholecystectomy				
. No	154(61.6)	79(61.7)	75(61.5)	*0.927*
. Yes	96(38.4)	49(38.3)	47(38.5)	
Previous Hernia Repair				
. No	59(24.6)	40(31.1)	19(15.6)	***0.005****
. Yes	191(76.4)	88(68.9)	103(84.4)	
Other				
. No	81(32.4)	38(29.7)	43(35.2)	*0.422*
. Yes	169(67.6)	90(70.3)	79(64.8)	
C-Section				
. No	226(91.4)	121(94.5)	105(86.1)	***0.040****
. Yes	24(9.6)	7(5.5)	17(13.9)	
Appendectomy				

. No	192(76.8)	105(82.0)	87(72.3)	*0.063*
. Yes	58(23.2)	23(18.0)	35(28.7)	
Ostomy				
. No	237(94.8)	120(94.6)	117 (95.9)	*0.631*
. Yes	13(5.2)	8(6.3)	5(4.1)	
Small Bowel Resection				
. No	246(98.4)	127(99.0)	119(97.5)	*0.581*
. Yes	4(1.6)	1(1.0)	3(2.5)	
Urological				
. No	249(99.6)	127(99.0)	122(100.0)	*0.981*
. Yes	1(0.4)	1(1.0)	0(0.0)	
Exploratory Laparotomy				
. No	237(94.8)	121(94.5)	116(95.1)	*0.929*
. Yes	13(5.2)	7(5.5)	6(4.9)	
# of Previous Hernias	2.0 [1.0, 3.0]	2.0 [1.0, 3.0]	2.0 [1.0, 3.0]	***0.044****
--Prev. Surgery Complications--				
Mesh Infection				
. No	227(90.8)	116(90.6)	111(91.0)	*0.903*
. Yes	23(9.2)	12(9.4)	11(9.0)	
Seroma				
. No	245(98.0)	127(99.0)	118(96.7)	*0.204*
. Yes	5(2.0)	1(1.0)	4(3.3)	
Other Complications				
. No	246(98.4)	125(97.7)	121(99.0)	*0.622*
. Yes	4(1.6)	3(2.3)	1(1.0)	
Wound Infection (6)				
. No	246(98.4)	125(97.7)	121(99.0)	*0.622*
. Yes	4(1.6)	3(2.3)	1(1.0)	
Procedure Hernia Mesh Size	900.0 [600.0, 1050.0]	900.0 [500.0, 1050.0]	930.0 [600.0, 1068.0]	*0.269*

Preoperative CCS Pain Score	1.60 [0.63-2.86]	1.00[0.25-2.31]	2.43[1.13-3.50]	***0.094***

**Denotes variables were statistically significantly different at p<0.05 between CCS score patient groups*

2.3 Results

Of the 250 patients who underwent abdominal wall reconstruction, 122(~48.8%) had a CCS score that was greater than 2. Of the 774 features, 710 were removed due to data missingness (features containing > 50% null values), <5% positive cases, or factors that were not deemed as preclinical variables. This filtering process left 64 independent variables for further analysis. Following this process, the remaining variables were assessed by Pearson's correlation coefficient for any notable variables with correlations >0.50 for multicollinearity. After assessing correlations, this left 57 candidate preclinical variables for analysis. Full patient characteristics for both experimental (CCS Score ≥ 2) and reference group (CCS Score < 2) for all preclinical variables are summarized in Table 2.1.

The patient cohort had a median age of 61 years old with an interquartile (IQR) range of 17 years. Approximately 42.4% of patients were male with the majority of patients in the assessed cohort being predominantly white (92.0%). The average patient in the population had an obese body mass index (BMI) score of >30.0 (median BMI was reported as 31.6 with an IQR of 7.2 units). In addition, patients had multiple existing medications, co-morbidities, and previous surgical histories. Between the two CCS score outcome groups, ten features were noted to be statistically significantly different including: patient gender, patient BMI, current tobacco usage, history of asthma, history of chronic pulmonary obstructive disease, history of diabetes, previous hysterectomy surgery, previous hernia repair, previous C-section, and number of previous hernia repairs.

As previously mentioned, the final multivariate model was selected after stepwise elimination of variables alongside multiple imputation, assuming the variables had no clinical or statistical importance for inclusion in the final model iteration. The complete multivariate model with adjusted odds ratios, confidence intervals, and unadjusted odds ratios is summarized in Table 2.2. The final model iteration was chosen based on statistical significance, clinical significance, as well as adjusted R-squared value. This model had 20 unique preclinical variables with an adjusted R-squared value of 0.2882. Of the 20 preclinical variables included in the model, four were significantly associated with the outcome with a P-value <0.05. Patients who are diabetic had OR=2.169 (95% CI: 1.032-4.561) of having a CCS score of ≥ 2 as compared to patients who are not diabetic, controlling for all other variables in the multivariate model. Patients who had a previous history of a colectomy surgery had OR=2.003 (95% CI: 1.009-3.975) of having a CCS score of ≥ 2 as compared to patients who did not have a previous history of colectomy surgery, controlling for all other variables. Similarly, patients who had a previous history of an appendectomy procedure had OR=2.205 (95% CI: 1.036-4.692) of having a CCS score of ≥ 2 as compared to patients who did not have a previous history of an appendectomy procedure. Finally, for every unit increase in patient preoperative CCS score, patients had OR=1.879 (95% CI: 1.441-2.450) of having a CCS score ≥ 2.

After listwise deletion of missing values, this left n=162 patients to generate the complete-case comparison data table. These data were also fitted using a logistic regression model. Of the 20 unique preclinical variables, four were statistically significantly associated with the outcome variable with a P-value of 0.05. Like the imputed model, the independent variables: patient diabetes status and preoperative CCS score were statistically significant indicating further evidence of the associations between these variables and patient QOL outcome. Of note, two

new variables were recorded as statistically significant with the outcome variable in the non-imputed dataset namely: patient history of hysterectomy procedure, and patient history of cholecystectomy procedure. It is possible these associations are becoming masked by the imputation technique, therefore obtaining more preoperative patient data may assist in determining the true relationship between these variables and the QOL outcome.

Table 2.2: Adjusted and unadjusted odds ratios for variables modeled by logistic regression

Variable	**Adjusted Odds Ratio [95% Confidence Interval]**	**Unadjusted Odds Ratio [95% Confidence Interval]**
Patient Age	0.991[0.968-1.014]	0.999[0.995-1.003]
Patient Sex		
. Female	Ref	Ref
. Male	0.670[0.306-1.464]	0.574[0.387-0.850]
Asthma		
. No	Ref	Ref
. Yes	1.617[0.557-4.688]	2.250[0.978-5.175]
Diabetes		
. No	Ref	Ref
. Yes	2.169[1.032-4.561]	1.565[0.928-2.641]
History of Cancer		
. No	Ref	Ref
. Yes	0.645[0.313-1.328]	0.659[0.405-1.070]
Hypertension		
. No	Ref	Ref
. Yes	0.665[0.345-0.1.282]	0.787[0.559-1.106]
Tobacco Use		
. No	Ref	Ref
. Yes	1.590[0.580-4.358]	2.200[1.042-4.646]
Hyperlipidemia		
. No	Ref	Ref

. Yes	0.707[0.302-1.655]	0.760[0.419-1.380]
Patient BMI	1.012[0.971-1.054]	1.001[0.993-1.008]
Previous Intra-abdominal Surgery or Trauma		
. No	Ref	Ref
. Yes	0.273[0.055-1.366]	0.952[0.741-1.224]
Hysterectomy		
. No	Ref	Ref
. Yes	1.527[0.666-3.502]	1.667[1.034-2.686]
Colectomy		
. No	Ref	Ref
. Yes	2.003[1.009-3.975]	1.15[0.724-1.817]
Cholecystectomy		
. No	Ref	Ref
. Yes	0.566[0.292-1.097]	0.959 [0.643-1.431]
Previous Hernia Repair		
. No	Ref	Ref
. Yes	1.891[0.899-3.976]	2.429[1.008-5.856]
C-Section		
. No	Ref	Ref
. Yes	2.583[0.812-8.219]	2.265 [1.056-4.859]
Appendectomy		
. No	Ref	Ref
. Yes	2.205[1.036-4.692]	1.522[0.899-2.575]
Race		
. White	Ref	Ref
. Non-White	0.693[0.228-2.111]	1.500[0.613-3.670]
Procedure Date		
. 2005-2011	Ref	Ref

. 2012-2017	0.678[0.359-1.281]	0.853 [0.611-1.191]
Mesh Infection		
. No	Ref	Ref
. Yes	0.522[0.1780-1.533]	0.917[0.404-2.077]
Preoperative CCS average pain	1.879[1.441-2.450]	1.241[1.105-1.393]

Table 2.3: Adjusted and unadjusted odds ratios for variables modeled by logistic regression for non-imputed data

Variable	**Adjusted Odds Ratio [95% Confidence Interval]**	**Unadjusted Odds Ratio [95% Confidence Interval]**
Patient Age	0.976 [0.947-1.006]	0.997 [0.992-1.002]
Patient Sex		
. Female	Ref	Ref
. Male	0.842 [0.313-2.261]	0.479 [0.291-0.788]
Asthma		
. No	Ref	Ref
. Yes	3.924 [0.844-18.259]	2.000 [0.602-6.642]
Diabetes		
. No	Ref	Ref
. Yes	3.139 [1.165-8.456]	1.533 [0.800-2.934]
History of Cancer		
. No	Ref	Ref
. Yes	0.516 [0.203-1.312]	0.484 [0.261-0.896]
Hypertension		
. No	Ref	Ref
. Yes	0.755 [0.314-1.815]	0.736 [0.687-1.112]
Tobacco Use		
. No	Ref	Ref
. Yes	1.575 [0.445-5.573]	2.143 [0.874-5.255]
Hyperlipidemia		

. No	Ref	Ref
. Yes	0.585 [0.188-1.819]	0.579 [0.276-1.216]
Patient BMI	1.021 [0.971-1.074]	0.998 [0.989-1.007]
Previous Intra-abdominal Surgery or Trauma		
. No	Ref	Ref
. Yes	0.508 [0.086-2.997]	0.857 [0.626-1.174]
Hysterectomy		
. No	Ref	Ref
. Yes	3.816 [1.220-11.934]	1.875 [1.022-3.439]
Colectomy		
. No	Ref	Ref
. Yes	1.803 [0.729-4.461]	0.917 [0.514-1.635]
Cholecystectomy		
. No	Ref	Ref
. Yes	0.281 [0.112-0.703]	0.765 [0.459-1.274]
Previous Hernia Repair		
. No	Ref	Ref
. Yes	2.365 [0.937-5.973]	1.105 [0.772-1.581]
C-Section		
. No	Ref	Ref
. Yes	2.383 [0.516-11.009]	2.750 [0.876-8.636]
Appendectomy		
. No	Ref	Ref
. Yes	2.015 [0.775-5.240]	1.352 [0.723-2.532]
Race		
. White	Ref	Ref
. Non-White	0.727 [0.167-3.152]	1.200 [0.366-3.932]

Procedure Date		
. 2005-2011	Ref	Ref
. 2012-2017	0.589 [0.254-1.364]	0.702 [0.468-1.051]
Mesh Infection		
. No	Ref	Ref
. Yes	1.611 [0.338-7.676]	1.750 [0.512-5.978]
Preoperative CCS average pain	1.436 [1.055-1.956]	1.151 [1.002-1.324]

After developing the multivariate model, the patient cohort sample was divided into a training and test sample with an 80:20 split. The test sample predictions were calculated using the established multivariate model. Model accuracy on the test sample was assessed through receiver operating characteristic (ROC) and was calculated at 0.847. To provide further validity to the calculated ROC score, 5-fold internal cross validation using stratified *k*-fold techniques was performed on the sample set using the same multivariate model. In this case, across the 5 generated test samples, the mean ROC was calculated at 0.750. Stratified *k*-fold sampling was chosen for incorporation to ensure the class frequency in the outcome label were preserved across test samples generated by the technique (Kluver et. al., 2016). ROC graph from 5-fold internal cross validation results are included in Figure 2.1 below.

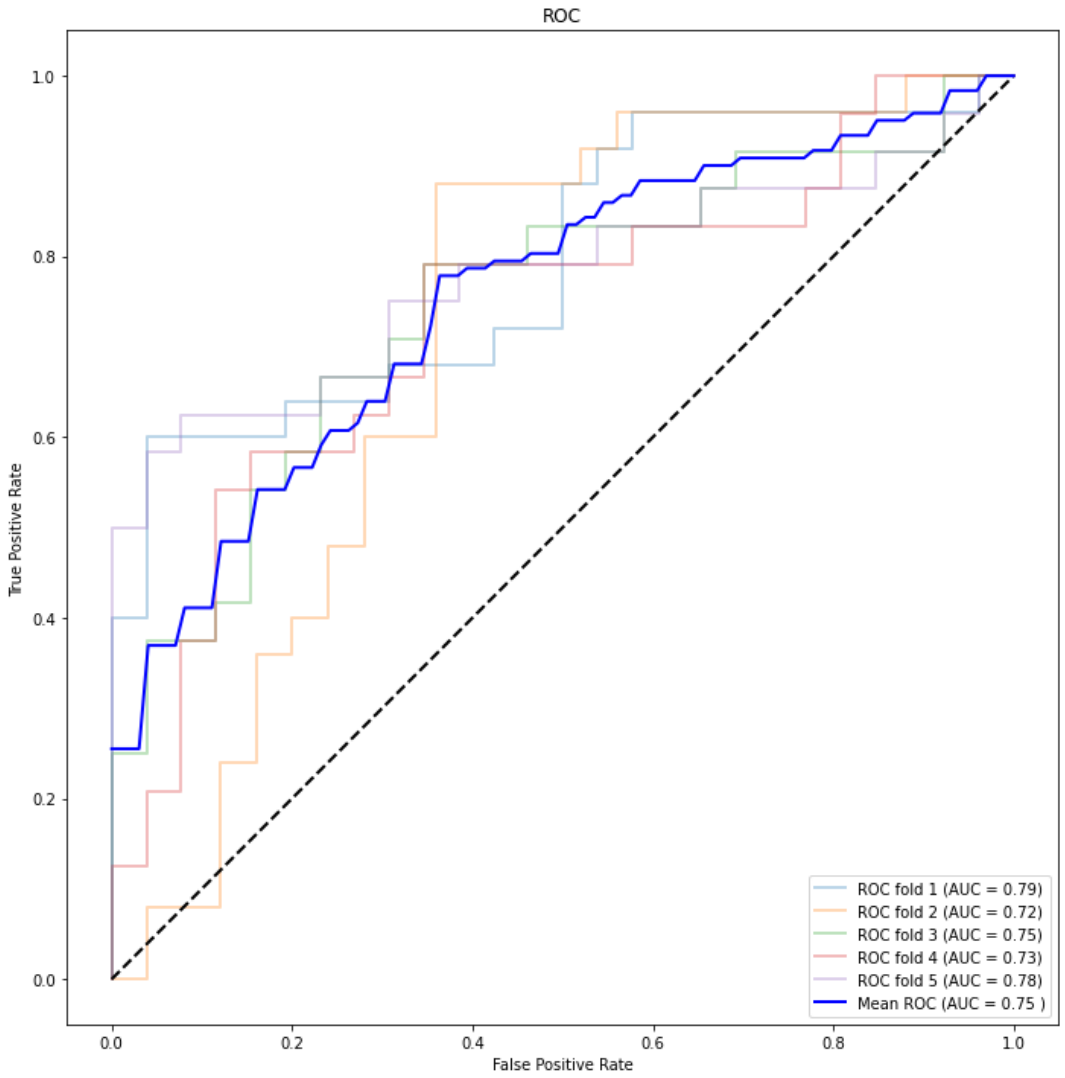


Figure 2.1: ROC curves from multivariate model predictions on 5-fold internal cross validation

2.4 Discussion

From the analysis of 250 patients who underwent abdominal wall reconstruction surgery (AWR) between 2005-2017, we developed a multivariable model to calculate the relationship of preoperative patient variables with patient QOL 6 months following surgery. Statistically significant predictors for reduced QOL following AWR surgery included patient age, patients taking coumadin, patients with diabetes, preoperative CCS scores, and prior history of hysterectomy surgery. Although not statistically significant, other preclinical variables were included in the model due to their clinical relevance to the QOL outcome. This included patient

sex, BMI, race, patients with end stage renal disease, asthma, hyperlipidemia, patient creatinine levels, and previous hernia surgery repair. Previous research has documented that patient age, sex, race, and BMI were all significantly associated with chronic pain, which is an indicator of QOL, following AWR surgery.

Some study strengths include utilizing the CCS Score to determine patient QOL. As mentioned previously, multiple validation studies have been performed by clinical institutions across the United States and internationally in the past decade (Cox et. al., 2016) and were documented to be preferred to the standard analog of the SF-36 survey. With prospective appraisals including 11,000 surveys, the instrument was shown to have significant acceptability and reliability for assessing QOL in patients undergoing hernia surgery (Heniford et. al., 2018). Another strength includes the relatively balanced classes among patients in the sample set between those with a CCS score ≥ 2 and those with a score <2. Due to a small sample set, having significant class imbalances would have made it extremely difficult to develop a predictive model to determine patient QOL 6 months following surgery.

Some limitations to the study include the relatively small sample size for developing an accurate and well-validated prediction model. There is an increased risk of model overfitting and therefore could have decreased generalizability to other sample populations. One possible solution to this issue is utilizing Firth's Logistic Regression corrections for small datasets which will be further assessed in future model analyses (Puhr et. al., 2017). Further investigation is therefore necessary to build a more robust prediction model with a larger sample set for internal validation to ensure more valid QOL predictions. In addition, another possible limitation could arise from the imputation techniques performed in data preprocessing. Although the MICE technique is a well-established and commonly applied technique for performing imputation on

missing data, these substituted values still hold risk of misrepresenting the true values in the dataset (Mera-Gaona et. al., 2021). Finally, since the study is dependent on patients submitting a 6-month follow-up survey, there is the possibility of loss to follow-up which reduces the number of potential participants to predict in the sample. Furthermore, previous studies have demonstrated through concepts like the Hawthorne effect that those participants who respond to surveys generally do not represent the average individual within the true population (Granberg & Holmberg, 1992). These problems can also be potentially addressed with larger sample sets in future analyses.

In conclusion, we developed a multivariate model based on 20 unique preclinical features to model the relationship between these variables and the outcome of interest: patient QOL 6 months following hernia surgery. This study furthers the research in developing a prediction model based purely on preoperative variables to determine QOL after surgery. Using these findings, we can continue to determine what preoperative factors reliably predict patient QOL 6 months after hernia surgery. Furthermore, we used this multivariate model to calculate predictions for patient QOL at 6 months using stratified k-fold cross validation techniques. The developed predictive model can assist clinicians during preoperative assessments to optimize early postoperative interventions and treatment plans for patients.

CHAPTER 2 REFERENCE LIST

Assunção, J. H., Malavolta, E. A., Gracitelli, M. E. C., Hiraga, D. Y., da Silva, F. R., & Ferreira Neto, A. A. (2017). Clinical outcomes of arthroscopic rotator cuff repair: correlation between the University of California, Los Angeles and American Shoulder and Elbow Surgeons scores. *Journal of Shoulder and Elbow Surgery*, 26(7), 1137–1142. https://doi.org/10.1016/j.jse.2017.01.025

Borad, N. P., & Merchant, A. M. (2017). The effect of smoking on surgical outcomes in ventral hernia repair: a propensity score matched analysis of the National Surgical Quality Improvement Program data. *Hernia: the Journal of Hernias and Abdominal Wall Surgery*, 21(6), 855–867. https://doi.org/10.1007/s10029-017-1664-1

Cox, T. C., Huntington, C. R., Blair, L. J., Prasad, T., Lincourt, A. E., Heniford, B. T., & Augenstein, V. A. (2016). Predictive modeling for chronic pain after ventral hernia repair. *The American Journal of Surgery*, 212(3), 501–510. https://doi.org/10.1016/j.amjsurg.2016.02.021

Du, J., Boss, J., Han, P., Beesley, L. J., Goutman, S. A., Batterman, S., Feldman, E. L., & Mukherjee, B. (2020). Variable selection with multiply-imputed datasets: choosing between stacked and grouped methods. https://doi.org/10.48550/arxiv.2003.07398

Granberg, D., & Holmberg, S. (1992). The Hawthorne Effect in Election Studies: The Impact of Survey Participation on Voting. *British Journal of Political Science*, 22(2), 240–247. https://doi.org/10.1017/S0007123400006359

Guyatt, G. H., Feeny, D. H., & Patrick, D. L. (1993). Measuring health-related quality of life. *Annals of internal medicine*, 118(8), 622–629. https://doi.org/10.7326/0003-4819-118-8-199304150-00009

Heniford, B. T., Walters, A. L., Lincourt, A. E., Novitsky, Y. W., Hope, W. W., &

Kercher, K. W. (2008). Comparison of generic versus specific quality-of-life scales for mesh hernia repairs. *Journal of the American College of Surgeons*, 206(4), 638–644.

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied Logistic Regression (3. Aufl.). Wiley.

Heniford, B. T., Lincourt, A. E., Walters, A. L., Colavita, P. D., Belyansky, I., Kercher, K. W., Sing, R. F., & Augenstein, V. A. (2018). Carolinas Comfort Scale as a Measure of Hernia Repair Quality of Life: A Reappraisal Utilizing 3788 International Patients. *Annals of surgery*, 267(1), 171–176. https://doi.org/10.1097/SLA.0000000000002027

Kehlet, H. (2008). Chronic pain after groin hernia repair. *British Journal of Surgery*, 95(2), 135–136. https://doi.org/10.1002/bjs.6111

Kluver, D., Mote, T., Leathers, D., Henderson, G. R., Chan, W., & Robinson, D. A. (2016). Creation and Validation of a Comprehensive 1° by 1° Daily Gridded North American Dataset for 1900–2009: Snowfall. *Journal of Atmospheric and Oceanic Technology*, 33(5), 857–871. https://doi.org/10.1175/JTECH-D-15-0027.1

Liang, M. K., Goodenough, C. J., Martindale, R. G., Roth, J. S., & Kao, L. S. (2015). External Validation of the Ventral Hernia Risk Score for Prediction of Surgical Site Infections. *Surgical Infections*, *16*(1), 36–40. https://doi.org/10.1089/sur.2014.115

Mera-Gaona, M., Neumann, U., Vargas-Canas, R., & Lopez, D. M. (2021). Evaluating the impact of multivariate imputation by MICE in feature selection. *PloS One*, 16(7), e0254720–e0254720. https://doi.org/10.1371/journal.pone.0254720

Montes Pérez, A., Roca, G., Sabaté, S., Lao, J. I., Navarro i Cuartiellas, A., Cantillo, J., Canet, J., & GENDOLCAT Study Group. (2015). Genetic and Clinical Factors Associated with Chronic Postsurgical Pain after Hernia Repair, Hysterectomy, and Thoracotomy: A Two-year

Multicenter Cohort Study. https://doi.org/10.1097/ALN.0000000000000611

Parseliunas, A., Paskauskas, S., Simatoniene, V., Vaitekunas, J., & Venskutonis, D. (2022). Adaptation and validation of the Carolinas Comfort Scale: a questionnaire-based cross-sectional study. *Hernia : the Journal of Hernias and Abdominal Wall Surgery*, 26(3), 735–744. https://doi.org/10.1007/s10029-021-02399-4

Puhr, R., Heinze, G., Nold, M., Lusa, L., & Geroldinger, A. (2017). Firth's logistic regression with rare events: accurate effect estimates and predictions? *Statistics in Medicine*, 36(14), 2302–2317. https://doi.org/10.1002/sim.7273

RStudio Team (2019). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA URL: http://www.rstudio.com/

Rois, R., Ray, M., Rahman, A., & Roy, S. K. (2021). Prevalence and predicting factors of perceived stress among Bangladeshi university students using machine learning algorithms. *Journal of Health, Population and Nutrition*, 40(1), 50–50. https://doi.org/10.1186/s41043-021-00276-5

Rudmik, L., Soler, Z. M., Mace, J. C., DeConde, A. S., Schlosser, R. J., & Smith, T. L. (2015). Using preoperative SNOT-22 score to inform patient decision for Endoscopic sinus surgery: Informing Shared Decision Making for ESS. *The Laryngoscope*, 125(7), 1517–1522. https://doi.org/10.1002/lary.2510

Schug, S. A., & Bruce, J. (2017). Risk stratification for the development of chronic postsurgical pain. *Pain Reports*, 2(6), e627–e627. https://doi.org/10.1097/PR9.0000000000000627

Urbach, D. R. (2005). Measuring Quality of Life After Surgery. Surgical Innovation, 12(2), 161–165. https://doi.org/10.1177/155335060501200216

Vad, M. V., Svendsen, S. W., Frost, P., Nattino, G., Rosenberg, J., & Lemeshow, S. (2022). Inguinal hernia repair among men: development and validation of a preoperative risk score for persistent postoperative pain. *Hernia: the Journal of Hernias and Abdominal Wall Surgery*, 26(1), 177–187. https://doi.org/10.1007/s10029-021-02376-x

van Ramshorst, G. H., Eker, H. H., Hop, W. C. J., Jeekel, J., & Lange, J. F. (2012). Impact of incisional hernia on health-related quality of life and body image: a prospective cohort study. *The American Journal of Surgery*, 204(2), 144–150. https://doi.org/10.1016/j.amjsurg.2012.01.012

Van Rossum, G., & Drake, F. L. (2009). Python 3 Reference Manual. Scotts Valley, CA: CreateSpace

Velanovich, V. (1998). Comparison of generic (SF-36) vs. disease-specific (GERD-HRQL) quality-of-life scales for gastroesophageal reflux disease. *Journal of Gastrointestinal Surgery*, 2(2), 141–145. https://doi.org/10.1016/S1091-255X(98)80004-8

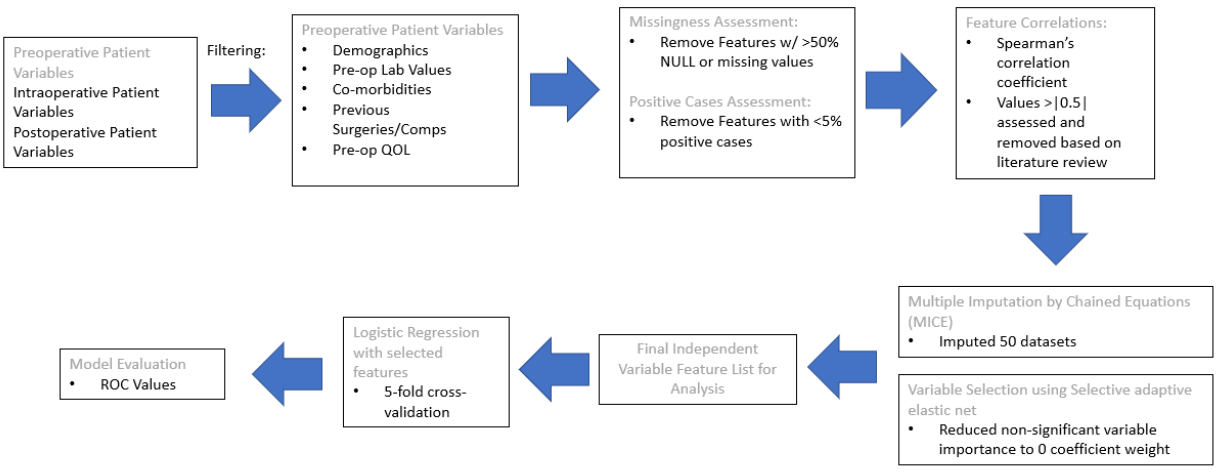
Wennberg J. E. (2004). Practice variation: implications for our health care system. Managed care (Langhorne, Pa.), 13(9 Suppl), 3–7.

APPENDIX A: UTILIZED SOFTWARE PACKAGES

Table denoting the software packages utilized in the analysis and their respective versions

Package	Version
Python (ver. 3.7.1)	
Pandas	1.1.1
Numpy	1.19.1
Scipy	1.5.2
Sklearn	1.0.2
Statsmodels	0.13.2
Matplotlib	3.3.1
Seaborn	0.11.2
Miceforest	3.1.1
R (ver. 4.0.0)	
glmer4	1.1-29

APPENDIX B: DIAGRAM OF DATA PROCESSING AND MODEL ANALYTICS



CHAPTER 3: OPTIMIZING IMAGE QUALITY FOR DEEP LEARNING CLASSIFICATION MODELS IN HERNIA SURGERY

3.1 Introduction

Developing accurate deep learning models based on medical imaging for surgery can be a daunting task; many deep learning models require significantly large training sample sizes of up to 10,000 images equally distributed among the classes for proper model development (Akkus et. al., 2017). Often, research questions in surgery deal with rare disease states and complications which provide particularly skewed datasets with outcomes of interest representing only 1%-5% of sampled individuals. These skewed datasets are particularly difficult for classification algorithms to sensitive results as it is very easy for the model to simply predict all samples as the negative or control group due to their extremely high prevalence in the dataset (LeMaitre, Fernando, & Aridas, 2017). Other potential limitations with utilizing clinical data include legal concerns and agreements, cost of data acquisition, and intensive labeling processes (Shorten et. al., 2019). Due to their low dataset frequency, researchers have therefore looked to augment these images either through artificially increasing their sample size in the dataset through image manipulation (a quantity-driven approach) or by modifying image characteristics to provide for more optimal classification by the deep learning model (a quality-driven approach) (Shorten et. al., 2019, Uemera et. al., 2021).

Data augmentation has traditionally been the most common method employed to combat these data limitation restrictions with clinical data (Lo et. al., 2021). Lo et. al. mention in their review that randomized data augmentation has represented the most popular form of the technique; however, this process generally does not accurately generate the natural randomness and variability necessary in image inputs to improve network robustness in classification. These random augmentations can include such techniques as flipping the image on a specific axis,

rotating the image, or modifying image characteristics which can potentially be too drastic or even too subtle to provide any appreciable improvements in network classification ability.

Traditionally in medical imaging, lighting conditions can vary widely across computed tomography (CT) scans and images due to common histopathological characteristics. Some examples of this would include the presence of air bubbles in the gastrointestinal tract, as well as metal implants (Machida et. al., 2010). These can cause surges in light intensity, or conversely appear as “blacked out” regions in the image. This variability or “noise” generated by these phenomena makes it difficult to classify any “regions of interest” present on the image and therefore finding optimal methods to reduce this noise is imperative to successful image classification.

With the desire to improve on classifying regions of interest on an image, many researchers have turned to utilizing techniques to first segment these regions prior to classification. One such method to improve classification is adaptive thresholding; this involves the binarization of an image (converting image pixels to 0 or 255) by setting a fixed threshold value (Balaji & Sumathi, 2014). Any pixel intensities less than the threshold value are consequently set to 0 while those above the threshold are set to 255. However, computing this optimal threshold value for an image can prove difficult and can require multiple iterations of hyperparameter tuning (Liu et. al., 2020). To address this limitation, researchers have utilized adaptive thresholding which considers small pixel clusters and computes the optimal threshold for that cluster. All adaptive and local thresholding methods assume that smaller regions of an image are more likely to have approximately uniform illumination (Liu et. al., 2020). This implies that local regions of an image will have similar lighting, as opposed to the image as a whole, which may have dramatically different lighting for each region.

In the context of image classification using patient surgical imaging, many patients undergo magnetic resonance imaging (MRI) or CT scans to identify the region of interest for operation. For example, the CT scan process takes detailed imagery of a patient's bones, tissues, organs, and muscles across various dimensional axes. The CT imaging software could have as few as a single image up to hundreds of image slices detailing the region of interest (Elhage et. al., 2019). As a result, a full set of image slices from an individual patient could detail multiple organ structures, bones, and muscles across multiple slices. This high level of variability among images within a patient poses another layer of complexity in classification, as an axial image taken at a patient's waistline will display very different structures compared to an image slice taken axially through the top of the chest even though both may pertain to the same abdominal outcome. Identifying methods of image optimization to reduce this variance is imperative to building successful classification models in surgery given the dearth of available data.

One possibility to circumvent some of this variance is utilizing a technique known as image blending. Here, an image composite is generated from a set number of input images of equal dimension. The pixel values at each location are averaged together to produce the resulting image, comprising equal parts of each input image. By averaging a set of images together, this aids in reducing some of the variability among a predetermined set of images, as the most frequent portions or characteristics are going to be the most represented in the image. Methods like image averaging could assist in removing the noise and variability present in CT imaging to provide more accurate predictions in image classification.

Although not the primary focus of this discussion, it is important to address the modeling architecture and methodology, as these modeling frameworks are ultimately used to classify these processed or unprocessed image sample sets. "ResNet", short for residual network, is a

type of deep learning neural network originally built in 2015 by He et. al. in image recognition competitions. The original 34-layer architecture won first place in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2015 classification competition with a top-5 error rate of 3.57% (An ensemble model). The researchers would go on to derive other model architectures including ResNet 18, ResNet-101, and ResNet-152 (He. Et. al., 2019). These models have been highly successful in image recognition and classification tasks across multiple publicly available datasets. Furthermore, clinical researchers have utilized these model architectures to great success, using the initial training weights garnered from previous datasets like ImageNet and applying them to answer multiple clinical research questions (Yu et. al., 2019, Chen et. al., 2022)

This study has two primary aims in investigating multiple qualitative approaches to improving image quality for classification in CT images. Traditionally, many deep learning image based studies utilize single images for building classification models (Lin et. al., 2021, Mittman et. al., 2022, Skrede et. al., 2020) The first aim will align and then average all CT images associated with a unique patient who underwent AWR surgery into a composite image to determine if the original sample or augmented sample produces a more accurate and discriminative dataset built on the ResNet-18 architecture. Model performance will be assessed on model accuracy and receiver operating characteristic. The second aim will test interactive adaptive thresholding techniques on all sample images and image averaging per patient to generate potential “optimized” experimental images. These experimental samples will also be incorporated into the ResNet-18 classification model and compared to the model utilizing the original surgery patient sample to determine any significant improvements in deep learning model classification accuracy and minimized loss function between the qualitatively improved patient samples and the original sample.

3.2 Methods

Patient Sample

This study incorporated data on n = 250 patients who underwent abdominal wall reconstruction (AWR) surgery between 2006-2017 at a tertiary care hernia referral center in the Eastern United States. Across all 250 patients, there were 4,225 CT scans available for analysis. In addition, patient demographic and comorbidity data were collected on the identified sample including patient age, gender, race, body mass index, hernia defect size, patient preoperative pain score, diabetes status, hypertension, and number of representative CT scans from a patient.

Image Preprocessing

Axial cut CT scans from the patient cohort were deidentified and rendered into representative 3-5mm slices using TeraRecon© software (TeraRecon Inc, Durham, NC) to ensure the sample abdominal CT images contained only the herniated abdominal region in the training set. CT image sizes after cropping to the region of interest were 512x512 pixels in length and width. All images were compiled per patient with identifiers removed. Finally, the processed output was extracted and stored in a secure folder for later AI model use in assessment and classification.

Prior to image classification or quantitative/qualitative optimization, all patient CT images were standardized to a size of 224x224 pixels. This input dimension size was chosen based on the necessary input dimensions for analysis by the ResNet-18 architecture. Images were also scaled to ensure that image aspect and resolution were preserved during standardization. Prior to data modification and model generation, the 250-patient sample was divided into training and test samples (80:20 training: test split) to ensure that the same test set was utilized

between the original and augmented/optimized samples. Based on the total sample, there were 3,380 CT scans in the training sample and 845 CT images in the testing sample.

Classification Outcome

The outcome variable, which was utilized to build the deep learning classification model to determine any improvements in classification accuracy, utilized the Carolinas Comfort Scale (CCS) Score. This score represents a well-validated and tested metric to assess patient quality of life after surgery (Heniford et. al., 2008). The metric has been both nationally and internationally validated in multiple AWR patient cohorts through Cronbach's α (0.97), test-retest validity, concurrent validity, and Spearman's correlation coefficient (Heniford et. al., 2018). Patient CCS scores were calculated from the sample and dichotomized to scores of ≥2, or <2. This binary dichotomization was chosen based on a previous study by Cox and colleagues (2016) which outlined this dichotomization of CCS scores of 0 (no symptoms) or 1 (minimal and not bothersome) as "asymptomatic patients" while those with CCS scores of 2 (signified as mild symptoms, but bothersome) or higher as "symptomatic patients". The binary QOL outcome [0-1, >2] was used to classify the patient sample by the ResNet-18 architecture.

Data Augmentation (Quantitative Approach)

To conduct the data augmentation techniques, the training dataset was duplicated to preserve an "original" sample set which does not have any transformations for comparison. The data augmentation step included multiple image transformations to the newly generated duplicate sample set. Each image underwent three transformations. First, all training images in the experimental set were rotated by 10 degrees. Next, all initial training images in the experimental set were zoomed by 10%. Finally, all initial images were center cropped. In summary, each patient's CT images were augmented by three extra images. This resulted in the initial training

set containing the original images, plus 3 generated images with a specific transformation (rotation, zoom, or axis flip). Examples of image transformations for data augmentation are illustrated in Figure 3.1 below.

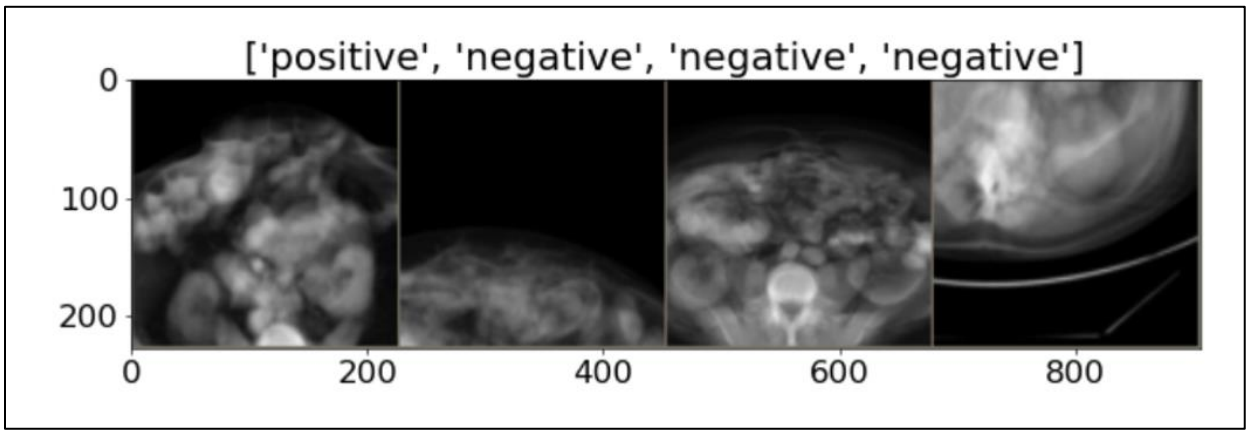


Figure 3.1: Example data augmentations from random images in the training sample. Images displayed show zoomed and cropped transformations of 4 example images.

Adaptive Thresholding (Qualitative Approach)

The adaptive thresholding was calculated and set using OpenCV's® adaptive thresholding function for localized regions in images (Kini et. al., 2021). The adaptive thresholding technique calculated the threshold values for pixels based on a specified region around a particular pixel. The original sample was therefore trained and compared to the patient image sample that underwent adaptive thresholding.

Image Averaging/Blending (Qualitative Approach)

Image averaging was conducted on each unique patient for his or her respective number of CT scans. A weighted average of the pixel values for the set of CT images with the same processed dimensions of 224x224 was calculated to produce the composite image. For example, a patient with 20 axial CT images was averaged together to produce a composite single image comprising the unique pixel values from those twenty images. The unmodified sample was trained and compared to the patient sample with image averaging to produce composites. An

example image set showing the unmodified image, the adaptive thresholding image, and the composite image average are included in Figure 3.2.

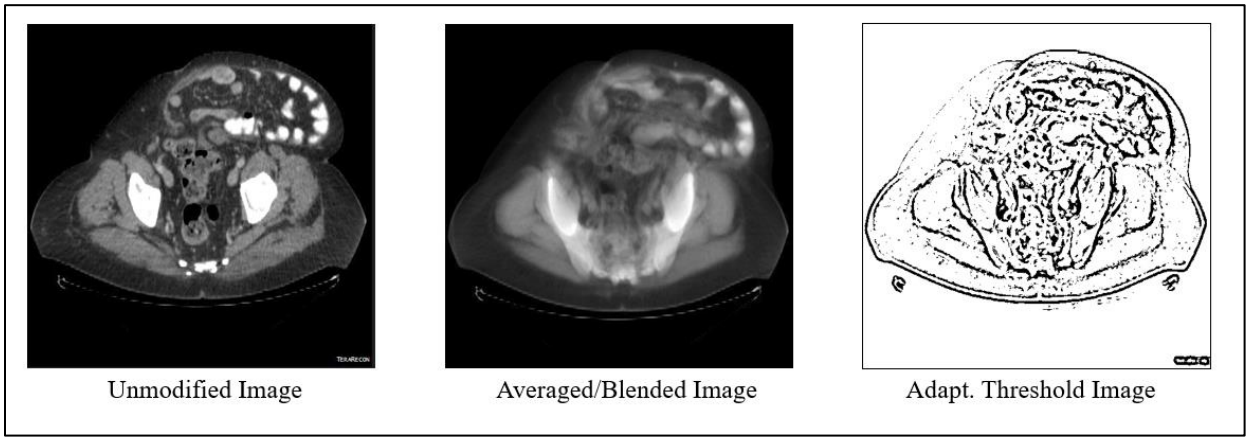


Figure 3.2: Image set (left to right) showing an example, unmodified single image from a patient, the composite image average for all images of that patient, and the adaptive threshold image.

Combining Qualitative Image Approaches

After assessing performance of both image blending and adaptive thresholding individually, both techniques were performed on the sample data. In this case, a patient with 20 axial CT images was averaged together to produce the composite image which then had adaptive thresholding applied to produce the final image product. This combined approach was trained and compared to the previous three approaches to determine which methodology had the greatest improvement in model accuracy and minimized loss function.

Model Generation

The deep learning model utilized for training was utilized for all discussed study samples. We utilized the established ResNET-18 architecture in PyTorch for model training and classification. The model was built using PyTorch software versions 1.13.1. The ResNet-18 model architecture consists of 18 unique layers including an initial convolutional layer, 4 sets of 4 convolutional layer sets, and a fully connected layer. The full model architecture and specifications are illustrated in detail in Appendix A. The ResNet18 model utilizes the stochastic

gradient descent (SGD) optimizer and sparse categorical cross-entropy loss function for model training (Geron, Aurelien, 2019). Transfer learning was employed using pretrained model weights for ResNET-18 on the ImageNet database. The architecture was used to model each dataset quantitative augmented only image sample set, quantitative augmented image sample set with adaptive thresholding, and quantitative augmented image sample set with composite image averages per patient. Model learning rate was set to 0.001 with 0.9 momentum. These models were trained for 100 epochs, and the highest model training accuracy from each respective trial run was recorded and saved for further analysis. Each model per unique dataset was run with the same model learning rate and momentum for consistency and comparison purposes. To provide further evidence towards results, models were cross validated with five trial runs per approach, and model training and validation accuracy were assessed per fold. Averaged results across all five runs were also calculated and reported.

Sample Set Evaluation

Statistical analysis for patient sample comparisons were conducted using Python version 3.7.1 (Van Rossum & Drake, 2009). Resnet model discriminative ability was assessed by percentage of correctly classified images on the test sample (Hosmer et. al., 2013). Evaluation metrics also included model loss calculated by binary cross entropy. First, the model trained on the original, un-augmented patient sample was compared by the outlined evaluation metrics to the model trained using the adaptive thresholding image sample. Next, the model trained on the original sample was compared by the same outlined evaluation metrics to the model trained using the images averaged per patient into composites. Finally, the model trained on a combination of composite/averaged images as well as adaptive thresholding was compared to

previous models to assess which qualitative improvements, if any, resulted in improved classification accuracy in the Resnet model.

3.3 Results

During the image processing steps, 33 patients were excluded from the analysis due to corrupted CT images. This left a total of n= 217 patients with varying numbers of CT images per patient for image augmentation and classification. Four unique datasets were generated and separately modeled using the ResNet-18 model architecture. These included: multiple images slices per patient with data augmentation (Model 1), multiple adaptive thresholding image slices per patient with data augmentation (Model 2), single composite/average image per patient with data augmentation (Model 3), and single composite/average adaptive thresholding image per patient with data augmentation (Model 4). Models 1 and 2 consisted of 4,799 training images and 1,173 testing images for an 80:20 train: test split for analysis. Models 3 and 4 consisted of 174 training images and 43 test images for analysis following the same 80:20 train: test split.

Each model was initially assessed through cross validation from 5 unique trial runs. Results for each model across each run, as well as the overall average training and validation accuracy and loss, are reported in tables 3.1-3.4. For Model 1 (multiple patient slices with data augmentation), the training accuracy was relatively high across all runs, with a reported average accuracy of 93.2736% (training loss: 0.1565). However, the validation accuracy in comparison was very low, averaging at 59.81244% (validation loss: 1.7351). Results were relatively consistent across trial runs, and there was minimal reported variance between runs (57.9710%-61.8926%). This likely indicates the model is overfitting on the training data and is unable to generalize well to the validation data.

Table 3.1: Cross-validation of 5 trial runs for model using single images (model 1) with reported training and validation accuracy and loss.

Trial	1	2	3	4	5	average
Epoch	9	18	26	90	10	30.6
Training Acc.	90.2271%	92.7902%	94.6864%	97.9579%	90.7064%	93.2736%
Training Loss	0.2285	0.1665	0.1270	0.0539	0.2068	0.1565
Val. Acc.	59.8465%	57.9710%	57.8858%	61.4663%	61.8926%	59.81244%
Val. Loss	1.1838	1.8999	1.9746	2.0810	1.5362	1.7351

Model 2 (multiple images per patient with adaptive thresholding) had relatively similar performance to model 1 with a slightly higher training accuracy (95.2448% vs. 93.2736%) with slightly lower training loss (0.1091 vs. 0.1565). However, model validation accuracy and validation loss were relatively comparable with only a slight improvement in accuracy observed in Model 2 over the five validation runs.

Table 3.2: Cross-validation of 5 trial runs for model using single images with adaptive thresholding (model 2) with reported training and validation accuracy and loss.

Trial	1	2	3	4	5	average
Epoch	34	79	67	92	40	62.4
Training Acc.	92.7902%	96.3951%	95.7074%	97.0827%	94.2488%	95.2448%
Training Loss	0.1526	0.0899	0.1028	0.0711	0.1293	0.1091
Val. Acc.	60.6991%	62.1483%	60.9548%	65.2174%	61.0401%	62.0119%
Val. Loss	1.8358	1.8536	2.0299	2.0697	1.8848	1.9348

Model 3 (average/composite image per patient) had relatively similar training accuracy. However, it excelled in validation accuracy and loss with the highest performance results compared to the two preceding models (74.4186% average validation accuracy). Some individual runs had reported validation accuracy greater than 80%, indicating good generalizability in Model 3 from the internal validation assessment.

Table 3.3**:** Cross-validation of 5 trial runs for model using composite image samples (model 3) with reported training and validation accuracy and loss.

Trial	1	2	3	4	5	average
Epoch	53	55	69	72	79	65.6
Training Acc.	94.8276%	91.9540%	89.6552%	95.4023%	93.1034%	92.9885%
Training Loss	0.1725	0.2074	0.1972	0.1094	0.1571	0.1687
Val. Acc.	74.4186%	69.7674%	76.7442%	69.7674%	81.3953%	74.4186%
Val. Loss	0.7551	0.9322	0.7744	0.9388	0.5964	0.7994

Finally, Model 4 (utilizing both average image and adaptive threshold techniques) showed no improvement in model accuracy nor loss function across any trial runs. Model validation loss was reportedly lower compared to Models 1 and 2. However, it did not exceed the performance reported from Model 3 across any trial run.

Table 3.4**:** Cross-validation of 5 trial runs for model using composite image samples with adaptive thresholding (model 4) with reported training and validation accuracy and loss.

Trial	1	2	3	4	5	average
Epoch	24	45	79	54	56	51.6
Training Acc.	84.8315%	91.5730%	89.8876%	93.8202%	92.1348%	90.4494%
Training Loss	0.3802	0.2817	0.1906	0.2081	0.2003	0.2522
Val. Acc.	58.6957%	63.0435%	63.0435%	56.5217%	60.8696%	60.4348%
Val. Loss	0.9852	1.1294	1.3547	1.4811	1.2359	1.2374

Average model test accuracies across trial runs, as well as minimized loss functions, are summarized in Table 3.5 below. The single image + adaptive thresholding with data augmentation (Model 2) had the highest observed training (approx. 96.6%) and lowest training loss (0.1091) compared to other models. However, all models did achieve a training accuracy greater than 90%. The average/composite image patient model (Model 3) had the highest observed training accuracy at 74.4186% and with the lowest corresponding validation loss compared to all other models (0.7994). Furthermore, no other model achieved a validation

accuracy >70%, suggesting model 3 as having the most generalizability among the four tested models.

Table 3.5: Averaged model training and test accuracies with corresponding loss function at the associated training epoch for all model types

	SINGLE IMAGE SAMPLE (MODEL 1)	**SINGLE IMAGE SAMPLE, ADAPT. THRESHOLD (MODEL 2)**	**AVERAGED IMAGE SAMPLE (MODEL 3)**	**AVERAGE IMAGE SAMPLE + ADAPT. THRESHOLD (MODEL 4)**
TRAINING ACCURACY	93.2736%	95.2448%	92.9885%	90.4494%
TRAINING LOSS	0.1565	0.1091	0.1687	0.2522
TEST ACCURACY	59.81244%	62.0119%	74.4186%	60.4348%
TEST LOSS	1.7351	1.9348	0.7994	1.2374
EPOCH	30.6	62.4	65.6	51.6

There was no observable improvement from utilizing adaptive thresholding on patient CT Images, and the calculated training and test accuracy was comparable to the base model (Model 1) without qualitative improvements. The combined model utilizing both adaptive thresholding and composite images had no major performance differences compared to both Models 1 and 2 (single image with augmentation, single image + adaptive thresholding with augmentation) and still had lower performance in accuracy and loss compared to Model 3 (composite/average image with augmentation). Of note, the validation loss in model 4 was observably lower than both Models 1 and 2. Overall, this suggests that averaging images per patient with data augmentation provided the greatest boost in training and test accuracy compared to other qualitative improvement techniques.

3.4 Discussion

In this study, we investigated two methodologies to qualitatively improve patient CT images with the goal of reducing image noise and variability and providing more standardized input data for model prediction and classification. Of the two techniques, generating composite images from the available CT scan slices available for a patient represented the greatest improvement in model validation accuracy and minimization of the associated loss function. This technique provides significant contributions to the existing literature surrounding image based classification models in surgery as many traditionally utilized single images without alignment for classification (Mittman et. al., 2022, Skrede et. al., 2020)

As mentioned in the introduction, patients may have multiple CT scan slices associated with their outcome, and with such high variability present between scans taken at varying positional axes in the chest, providing a composite average may assist in model discretization on the most important features pertaining to the classification outcome or outcomes. The ResNet18 model trained on these composite images outperformed all other models in both validation accuracy and loss, suggesting this technique as a promising means of improving image quality for classification of surgical outcomes.

In contrast, the adaptive thresholding technique did not demonstrate any improvements in model classification accuracy as compared to the baseline image set with no qualitative transformations. In addition, the image sample utilizing both adaptive thresholding and image averaging performed worse than the model utilizing only image averaging. This suggests that the adaptive thresholding technique may be removing or masking important feature information within the image, resulting in no improvements to the classification accuracy after modeling.

This research contributes to the growing body of literature surrounding deep learning classification models and their applications within the surgical field. Numerous studies have

already been published utilizing deep learning models for prediction using medical imaging such as CT scans. Within the realm of abdominal wall reconstruction surgery, Elhage et. al., 2019 successfully built a deep learning model to predict patient outcomes following abdominal wall reconstruction surgery with successful model performance. However, this study and others have not sought to investigate the image data inputs to effectively modify these inputs for superior classification. Due to the limited amount of available data for predictive modeling in surgery, this study addresses this fundamental issue by providing researchers the means to utilize as much available image data as possible without having to discard candidate medical images due to too much noise, image corruption or other issues. Within the context of AWR surgery, the image alignment and averaging techniques provide a standardized process for image input data to be effectively analyzed by predictive models and produce much higher performance when compared to using unaltered or unmodified input images.

Currently, no deep learning classification models exist for determining patient QOL using preoperative imaging data. Reduced patient QOL generally manifests as chronic pain postoperatively which also represents the most prevalent adverse outcome following hernia surgery. Effective management of chronic postoperative pain is a highly expensive process for patients with 9-year cost estimates exceeding $50,000 for U.S patients (Elsamadicy et. al., 2019). Classification models like those generated in this study can assist clinicians in identifying patients at reduced risk for QOL after surgery. Successful classification can therefore provide clinicians with the information to optimize treatment plans that are best suited to reducing financial costs while also optimizing patient health after surgery.

Although this study aimed to explore some qualitative methods of improving image quality in medical images, this is not an exhaustive analysis of available techniques. Recent

research from Nagayam et. al., (2018) utilized deep learning reconstruction (DLR) techniques to successfully reduce CT image noise and increase spatial resolution compared to standard image reconstruction techniques like filtered back projection. Similar to DLR, Chaurdari et. al. (2019) explored the use of "super-resolution" techniques which utilize convolutional neural networks to transform low resolution MRI images to higher order resolution for detection of osteoarthritis. Notably, both studies were not concerned with improving image quality for model classification, but rather to improve image quality to assist clinical decision making with more readily interpretable images.

Another possible study limitation is the utilization of a singular deep learning architecture with ResNet-18. With the field of artificial intelligence constantly evolving and new neural network architectures being built and becoming available under various frameworks, there is the possibility that some architecture may better fit and classify the input data than ResNet-18. For example, ResNet-18 was originally trained on ImageNet which comprises many categories of natural images which has stark differences in feature characteristics compared to medical images. Consequently, Alzubaidi and colleagues (2021) developed a DLM: "MedNet", which was specifically trained on over 3 million publicly available types of medical image data including CT scans. Future research could incorporate investigating multiple DLMs such as MedNet to compare performance across models in addition to image input optimization.

In conclusion, multiple qualitative methodologies to improve medical image quality for analysis by prediction models were assessed. There are many tools available to researchers in manipulating and augmenting surgical images, however, it is difficult to determine what techniques are going to be most optimal for improving image quality for better evaluation performance by deep learning models. Furthermore, this analysis is very context-specific, and

what may be applicable for surgical and medical imaging may not be useful in developing methodologies in other fields. Image averaging therefore represents a promising means of optimizing patient CT images for AWR surgery for classification. Utilizing techniques like image averaging aids researchers in maximizing the limited image data available for building classification models in medicine.

CHAPTER 3 REFERENCE LIST

Akkus, Z., Galimzianova, A., Hoogi, A., Rubin, D. L., & Erickson, B. J. (2017). Deep Learning for Brain MRI Segmentation: State of the Art and Future Directions. *Journal of Digital Imaging*, 30(4), 449–459. https://doi.org/10.1007/s10278-017-9983-4

Alzubaidi, L., Santamaría, J., Manoufali, M., Mohammed, B., Fadhel, M. A., Zhang, J., Al-Timemy, A. H., Al-Shamma, O., & Duan, Y. (2021). MedNet: Pre-trained Convolutional Neural Network Model for the Medical Imaging Tasks. https://doi.org/10.48550/arxiv.2110.06512

Balaji, T., & Sumathi, M. (2014). Effective Features of Remote Sensing Image Classification Using Interactive Adaptive Thresholding Method. arXiv.org.

Chaudhari, A. S., Stevens, K. J., Wood, J. P., Chakraborty, A. K., Gibbons, E. K., Fang, Z., Desai, A. D., Lee, J. H., Gold, G. E., & Hargreaves, B. A. (2020). Utility of deep learning super-resolution in the context of osteoarthritis MRI biomarkers. *Journal of Magnetic Resonance Imaging*, 51(3), 768–779. https://doi.org/10.1002/jmri.26872

Cox, T. C., Huntington, C. R., Blair, L. J., Prasad, T., Lincourt, A. E., Heniford, B. T., & Augenstein, V. A. (2016). Predictive modeling for chronic pain after ventral hernia repair. *The American Journal of Surgery*, 212(3), 501–510. https://doi.org/10.1016/j.amjsurg.2016.02.021

Elsamadicy, A. A., Ashraf, B., Ren, X., Sergesketter, A. R., Charalambous, L., Kemeny, H., Ejikeme, T., Yang, S., Pagadala, P., Parente, B., Xie, J., Pappas, T. N., & Lad, S. P. (2019). Prevalence and Cost Analysis of Chronic Pain After Hernia Repair: A Potential Alternative Approach With Neurostimulation. Neuromodulation (Malden, Mass.), 22(8), 960–969. https://doi.org/10.1111/ner.12871

Geron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. O'Reilly Media, Incorporated.

Heniford, B. T., Walters, A. L., Lincourt, A. E., Novitsky, Y. W., Hope, W. W., & Kercher, K. W. (2008). Comparison of generic versus specific quality-of-life scales for mesh hernia repairs. *Journal of the American College of Surgeons*, 206(4), 638–644. https://doi.org/10.1016/j.jamcollsurg.2007.11.025

Heniford, B. T., Lincourt, A. E., Walters, A. L., Colavita, P. D., Belyansky, I., Kercher, K. W., Sing, R. F., & Augenstein, V. A. (2018). Carolinas Comfort Scale as a Measure of Hernia Repair Quality of Life: A Reappraisal Utilizing 3788 International Patients. *Annals of surgery*, 267(1), 171–176. https://doi.org/10.1097/SLA.0000000000002027

Ho, D., Liang, E., Stoica, I., Abbeel, P., & Chen, X. (2019). Population Based Augmentation: Efficient Learning of Augmentation Policy Schedules. arXiv.org.

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied Logistic Regression (3. Aufl.). Wiley.

Hu, J., Yan, C., Liu, X., Li, Z., Ren, C., Zhang, J., Peng, D., & Yang, Y. (2021). An integrated classification model for incremental learning. Multimedia Tools and Applications, 80(11), 17275–17290. https://doi.org/10.1007/s11042-020-10070-w

Lemaitre, G., Nogueira, F., & Aridas, C. K. (2017). Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *Journal of Machine Learning Research*, 18, 1–5.

Lin, Q., Cao, C., Li, T., Man, Z., Cao, Y., & Wang, H. (2021). dSPIC: a deep SPECT image classification network for automated multi-disease, multi-lesion diagnosis. *BMC Medical Imaging*, 21(1), 122–122. https://doi.org/10.1186/s12880-021-00653-w

Lis, K., Korycinski, M., & Ciecierski, K. A. (2021). Classification of masked image data. PloS One, 16(7), e0254181–e0254181. https://doi.org/10.1371/journal.pone.0254181

Liu, L. Y., Liu, Y., & Zhu, H. (2020). Masked convolutional neural network for supervised learning problems. Stat (International Statistical Institute), 9(1). https://doi.org/10.1002/sta4.290

Lo, J., Cardinell, J., Costanzo, A., & Sussman, D. (2021). Medical Augmentation (Med-Aug) for Optimal Data Augmentation in Medical Deep Learning Networks. Sensors (Basel, Switzerland), 21(21), 7018–. https://doi.org/10.3390/s21217018

Machida, H., Masukawa, A., Tanaka, I., Fukui, R., Suzuki, K., Ueno, E., Kodera, K., Nakano, K., & Shen, Y. (2010). Prospective Electrocardiogram-Gated Axial 64-Detector Computed Tomographic Angiography vs Retrospective Gated Helical Technique to Assess Coronary Artery Bypass Graft Anastomosis: Comparison of Image Quality and Patient Radiation Dose. *Circulation Journal*, 74(4), 735–740. https://doi.org/10.1253/circj.CJ-09-0714

Mittmann, B. J., Braun, M., Runck, F., Schmitz, B., Tran, T. N., Yamlahi, A., Maier-Hein, L., & Franz, A. M. (2022). Deep learning-based classification of DSA image sequences of patients with acute ischemic stroke. *International Journal for Computer Assisted Radiology and Surgery*, 17(9), 1633–1641. https://doi.org/10.1007/s11548-022-02654-8

Kini, M. S., Bhandarkar, R., & Shenoy, K. P. (2021). Real Time Moving Vehicle Congestion Detection and Tracking using OpenCV. *Turkish Journal of Computer and Mathematics Education*, 12(10), 273–279.

Nagayama, Y., Sakabe, D., Goto, M., Emoto, T., Oda, S., Nakaura, T., Kidoh, M., Uetani, H., Funama, Y., & Hirai, T. (2021). Deep Learning-based Reconstruction for Lower-Dose Pediatric CT: Technical Principles, Image Characteristics, and Clinical Implementations. *Radiographics : a review publication of the Radiological Society of North America, Inc*, *41*(7), 1936–1953. https://doi.org/10.1148/rg.2021210105

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data*, 6(1), 1–48. https://doi.org/10.1186/s40537-019-0197-0

Skrede, O.-J., De Raedt, S., Kleppe, A., Hveem, T. S., Liestøl, K., Maddison, J., Askautrud, H. A., Pradhan, M., Nesheim, J. A., Albregtsen, F., Farstad, I. N., Domingo, E., Church, D. N., Nesbakken, A., Shepherd, N. A., Tomlinson, I., Kerr, R., Novelli, M., Kerr, D. J., & Danielsen, H. E. (2020). Deep learning for prediction of colorectal cancer outcome: a discovery and validation study. *The Lancet* (British Edition), 395(10221), 350–360. https://doi.org/10.1016/S0140-6736(19)32998-8

Uemura, T., Näppi, J. J., Ryu, Y., Watari, C., Kamiya, T., & Yoshida, H. (2021). A generative flow-based model for volumetric data augmentation in 3D deep learning for computed tomographic colonography. *International Journal for Computer Assisted Radiology and Surgery*, 16(1), 81–89. https://doi.org/10.1007/s11548-020-02275-z

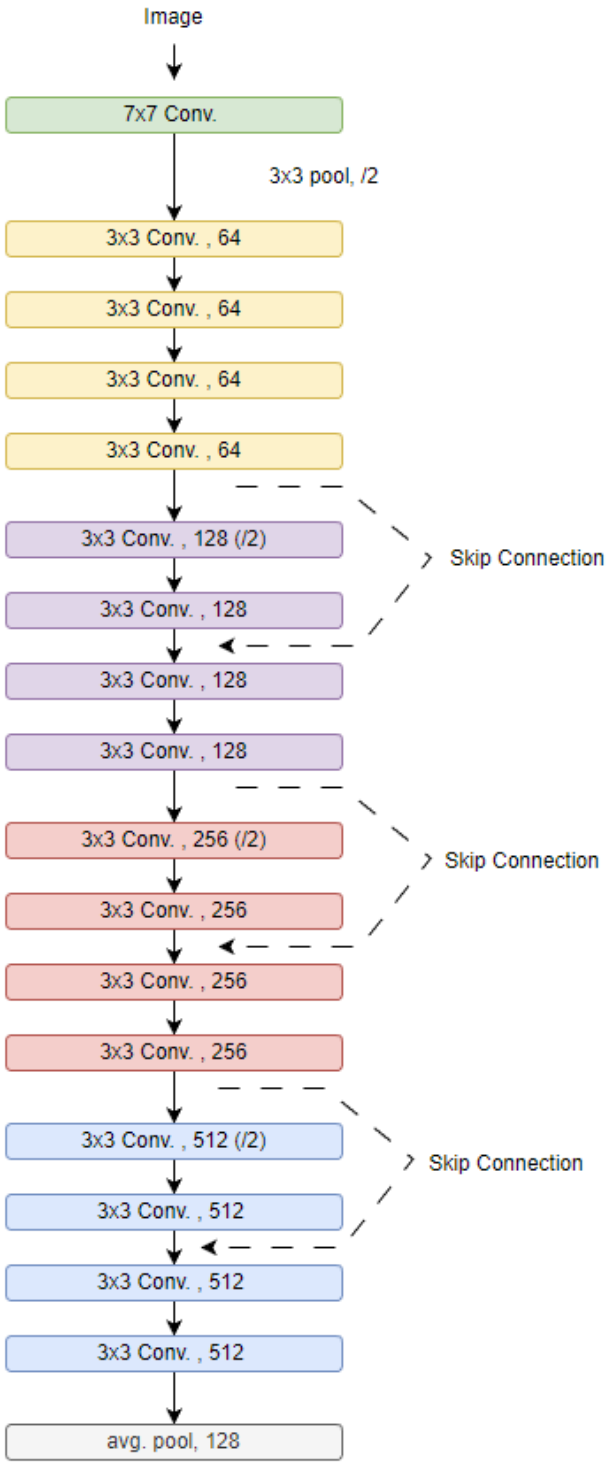
Van Rossum, G., & Drake, F. L. (2009). *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace.

Yaman, B., Hosseini, S. A. H., Moeller, S., & Akcakaya, M. (2021). Improved Supervised Training of Physics-Guided Deep Learning Image Reconstruction with Multi-Masking. ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 1150–1154. https://doi.org/10.1109/ICASSP39728.2021.9413495

Zhang, C., Cui, J., & Yang, B. (2019). Learning Optimal Data Augmentation Policies via Bayesian Optimization for Image Classification Tasks.

https://doi.org/10.48550/arxiv.1905.02610

APPENDIX C: RESNET-18 ARCHITECTURE DIAGRAM SHOWING ALL 18 LAYERS



CHAPTER 4: VALIDATING A NOVEL DEEP LEARNING MODEL FOR ABDOMINAL WALL RECONSTRUCTION WITH EXTERNAL COHORTS

4.1 Introduction

In recent decades, there has been an exponential increase in the number of surgical prediction models using artificial intelligence (AI) and deep learning (Chen et. al., 2018, Lee et. al., 2020). Deep learning systems employ an architectural framework consisting of artificial neural networks (ANN) which consist of multiple individual processing units commonly termed "neurons" that incorporate sets of input features with their corresponding weights (Savadijev et. al., 2019). The features and weights are summed as inputs and passed through a classification function to determine the output of the neuron. This methodology is intended to mimic a simplified architectural structure and function of a biological neuron which similarly incorporates multiple chemical inputs to send an output signal to other neurons (Bahl, M. 2020). Consequently, these hierarchical networks can encode complex nonlinear functions with significantly large numbers of input features or variables. For example, deep linear models have traditionally performed extremely well at classification problems involving images as the model contains the necessary structure that is directly sensitive to input data including the individual pixels in an image (Montagnon et. al., 2020). These pixels can subsequently be evaluated to determine potential classification outputs based on the proposed problems.

These models have shown great promise and high accuracy performances in predicting patient outcomes; however, with their low levels of interpretability when calculating predictions, these models require substantial scrutiny before proper clinical implementation (Savadijev et. al., 2019). Other researchers have noted that even slight modifications to the input data in deep learning models can result in dramatically different classifications, suggesting that these models must be extensively investigated and validated to ensure generalizability (Szegedy et. al., 2013).

These deep learning models utilize data-based algorithms to determine a patient's risk for a certain outcome of interest depending on patient characteristics (Ramspek et. al., 2021). Such models have gained increased interest from medical practitioners and researchers aiming to develop models capable of optimizing a patient's treatment plan by assessing benefit-to-risk ratios. Although this development has improved patient treatment plans and standardized care in some fields, the full potential of these models has yet to be seen (Navarro et. al., 2021). Many deep learning models were initially developed on a specific cohort of patients; very few were later externally validated, which is imperative to proving model reproducibility and generalizability to other cohorts (Collins et. al., 2015). Among the approximately 85,000 prediction model publications available on PubMed® in surgery, less than 5% included some form of external validation (Ramspek et. al., 2021). Since model performance is generally worse in external samples than in the initial sample used for model training, prediction models should not be recommended for clinical implementation and use before external validation (Fatemi et. al., 2021). Furthermore, it is imperative to appropriately apply external validation methodologies in clinical studies and also where it fails to guide us towards improved recommendations with clinical procedures.

Of primary concern to both researchers and clinicians is that the prediction model initially developed on a specific dataset from a patient population may show excellent performance and accuracy, however, these results may not translate to external cohorts (Siontis et. al., 2015). To determine model performance, multiple measures can subsequently be included to test for model discrimination, calibration, and accuracy on the new external cohorts (Collins et. al., 2015). Siontis and colleagues (2015) investigated 127 risk prediction models published on clinical decision-making on the Pubmed® biomedical literature database to ascertain what

validation metrics, if any, were incorporated into the respective publication's research design. They determined that only 25% of original authors/groups or third parties externally validated these initially published prediction models within five years from the initial model publication. In addition, the authors noted a statistically significant reduction in the mean area under the curve (AUC) score on models externally validated after initial publication (Siontis et. al., 2015).

Elhage et al (2021) recently developed a novel deep learning–based prediction model incorporating 9,303 computed tomography (CT) images from 369 patients undergoing abdominal wall reconstruction (AWR) surgery to determine the probability of a patient requiring a component separation procedure based solely on these preoperative imaging scans. The researchers incorporated supervised learning techniques whereby the neural network is provided with labeled images in the training sample to learn the defining features directly from these images without the need for additional information (Starke et. al., 2020). The initial study was successful in internally validating the model with an AUROC (Area Under the Receiver Operating Characteristic) of 0.744 predicting the chance of needing component separation technique (CST) in abdominal wall reconstruction patients based on their preoperative CT images. The researchers also trained a second model on predicting the chance of surgical site infection (SSI) within 30 days of surgery and reported an AUROC value of 0.898 from internal validation. However, the clinical validity and reproducibility of this deep-learning model remain unknown without further validation from external cohorts (Navarro et. al., 2021).

Originally, this study aimed to externally validate both previously developed deep learning models on an AWR patient cohort from a tertiary hernia center in the South Piedmont area of the United States (Elhage et. al., 2021). The initial assessment used the same model architecture and pre-trained model weights used in Elhage et. al's 2021 paper for external

validation using a cohort of 75 patients from the Ohio State University Medical School. However, the model performance on the external cohorts was extremely low with reported receiver operating characteristic (ROC) values of 0.51 and 0.49 for CST and SSI models, respectively.

As a result, this study sought to fit the original patient image data using a new deep-learning model architecture to model two separate outcomes: the probability of requiring a component separation procedure and the probability of developing a wound infection after AWR surgery within 30 days. These models would first be trained and internally validated, then externally validated using the previously mentioned cohort of patients from the Ohio State University Medical School. Model Evaluation was based on discriminative ability including evaluation by comparison of ROC value and overall classification accuracy with corresponding loss functions.

4.2 Methods

Patient Cohorts:

The initial patient sample consisted of n=362 candidates from Atrium Health Carolinas Medical Center who underwent open AWR surgery to model CST and SSI outcomes. All data (n=362) were available to model the probability of SSI outcome in the patient sample. Of the 362 candidates, CST outcome data were available for n=297 patients for model training to predict the probability of requiring CST. Both SSI and CST patient cohorts were utilized for initial model training and internal validation through leave-one-out cross-validation (LOOCV) and k-fold cross validation (Meijer & Goeman, 2013). Model external validation was conducted using a patient cohort consisting of 75 patients from the Ohio State Medical School comprising 75 preoperative CT scans. All patient cohorts will be identified retrospectively from existing

databases of patients who underwent AWR surgery. Patient outcomes in each cohort will be identified to determine if the patient required a CST. Those requiring a CST will be defined as a positive “yes” case if the patient underwent musculo-fascial advancement via either an external oblique release (EOR) or transversus abdominus release (TAR) to enable complete abdominal wall closure.

The SSI outcome was defined as either deep or superficial wound infection which was previously determined and documented by data entry specialists based on sources from the hospitals, clinics, and direct patient follow-ups, as well as any associated imaging documentation. Patient outcomes were dichotomized as either “0”, indicating no identified wounds/infections within 30 days of surgery, or “1”, indicating presented wounds/infections within 30 days of surgery. Individual image classification models were built for both CST and SSI outcomes and were utilized for sample assessment and evaluation following prediction scoring.

In addition, patient demographic data were collected on each cohort which included age, gender, race, body mass index, hernia defect size, comorbidities, and Centers for Disease Control and Prevention wound classification to characterize each sample (Yamamoto et. al., 2015).

Patient Inclusion/Exclusion Criteria

After receiving Institutional Review Board approval, patients were retrospectively identified from each respective institution from internally managed databases. Patients were excluded from the analysis if they were less than 18 years old, or had an emergent abdominal operation at the time of AWR. Patients were excluded if preoperative CT images displayed significant distortion or noise (for example, light saturation resulting from an orthopedic

prosthetic). In addition, patients with missing CT images around the region of interest (herniated region) are excluded from the analysis.

Image Preprocessing

Axial cut patient CT scans were deidentified and rendered into representative 3-5mm slices using TeraRecon© software (TeraRecon Inc, Durham, NC) so that only abdominal CT images containing the hernia region were included in the training set. All images were batched by patient, all patient identifiers were removed, and were exported and stored in a secure folder for later use by AI model assessment and validation.

Before deep learning model classification, images were standardized to a size of 224 x 224 pixels. Images were also scaled to ensure that image aspect and resolution are preserved during standardization. Since each patient had multiple associated CT image slices, each individual's CT images were "blended" by taking the average pixel value across *n* images where *n* is the number of unique CT image slices associated with a unique patient. We, therefore, generated a "composite" CT image per patient for input and analysis by classification model. The final image number for internal training and validation is 297 images for the Atrium sample CST outcome. For the SSI outcome, there were 362 composite images from Atrium patients available. The external Ohio validation samples for CST and SSI, consisted of 75 composite images from each patient for each outcome.

Model Generation

We utilized the established ResNET-18 architecture in PyTorch for model training and classification (Paszke et. al., 2019). This architecture was utilized for internal and external validation for both surgical outcomes (CST and SSI). The model was built using PyTorch software version 1.13.1. The ResNet-18 model architecture comprises 18 unique layers which

includes the initial convolutional layer, four sets of convolutional layers of similar filter size, and the final fully connected layer. The full model architecture and specifications are illustrated in detail in Appendix C. The ResNet-18 model utilizes the stochastic gradient descent (SGD) optimizer and sparse binary cross-entropy loss function for model training (Geron, 2019). Transfer learning was employed using pretrained model weights for ResNET-18 on the ImageNet database. For cross-validation, datasets fit to Resnet-18 were trained for 100 epochs with a predetermined batch size of 32. Early stopping was also employed to cease training when the loss function was not found to decrease any further after five subsequent epochs. Model weights and settings at this epoch were saved for later evaluation on the external validation set.

Leave-one-out Cross-validation

ResNet model classification consistency was first assessed by modeling patient outcomes using LOOCV (Meijer & Goeman, 2013). This technique was used in conjunction with k-fold cross-validation across multiple training runs to provide less biased estimates in internal prediction. Each image received an individual prediction classification using *n-1* images for training where *n* = the total sample of images employed in the study. This process would thus produce *n* models for each classification. Model predictions from LOOCV were recorded and the average classification accuracy across all patient samples was recorded for both the CST and SSI models separately.

Model Tuning and Cross-Validation

To ensure the ResNet-18 model had optimal fit on the patient sample data, multiple learning rate runs were tested from the specified list: [0.01, 0.001, 0.0001] which were used to determine the magnitude with which to update model weights during training (Iiuduka, 2022). In addition, for model training and internal validation, cross-validation across five model trial runs

was conducted to provide a more robust result on model accuracy during initial internal training and validation. Finally, sample sets with class imbalances between negative and positive outcomes for either CST or SSI were controlled for appropriately using the weighted class penalization function where incorrect classifications on the less frequent class are more heavily penalized during model training compared to the other.

Model Predictions

All patient test images were inputted into the existing deep learning model to generate classification scores for each outlined outcome. Classification percentages for all test images were therefore calculated from the pre-trained model weights. Each patient analyzed by the model received a continuous prediction value ranging from 0 to 1, with values >0.5 denoting a "positive" (1) outcome for either model, i.e. probability of CST or SSI. Scores less than the determined threshold are set as 0 or a "negative" outcome. These prediction scores were joined to the actual patient outcome label for subsequent evaluation. Each patient within each external cohort received a prediction score for the risk of requiring a component separation procedure, and the risk of postoperative SSI.

Model Evaluation

Statistical analysis for model evaluation was conducted using Python version 3.7.1 (Van Rossum & Drake, 2009). First, model accuracy was assessed by LOOCV to determine consistency in predictions. For the internal cross-validation, evaluation metrics also included model training and validation accuracy and corresponding loss functions. In addition, models were assessed by discriminative ability through the ROC score (Hosmer et. al., 2013). External validation similarly utilized both percentage of correctly classified outcomes and the corresponding ROC value.

4.3 Results

Cohort Description

For the SSI model, the internal sample (Atrium-Carolinas Medical Center) consisted of n = 362 patients where 285 were negative and 77 were positive. The CST model consisted of n = 297 patients of which 200 had a negative CST outcome and 97 had a positive outcome. The external Ohio cohort consisted of 75 patients where 62 were negative for SSI and 13 were positive. The CST outcome had 27 who were negative and 48 who were positive.

Leave-one-out Cross-validation

To build the deep learning models under the Resnet-18 architecture, the internal patient samples were first separated by outcome into two cohorts: CST and SSI. Model results from LOOCV are summarized in table 4.1 per outcome measure. The SSI model demonstrated promising classification performance with a reported accuracy of 94.65% from LOOCV across all 362 images. The CST model had reasonable overall classification accuracy with 75% of cases correctly classified from LOOCV. Although not exhaustive, reported accuracies from LOOCV provide an initial assessment of model internal performance with less biased estimates than simply utilizing a single train/test split.

Table 4.1: Leave-one-out cross validation accuracy results for component separation and surgical site infection models using internal patient sample.

	Prediction Accuracy (Full data)	Correctly Classified (LOOCV)
Surgical Site Infection	93.37%	92.29%
Component Separation	77.44%	75.00%

Internal Validation

For the SSI and CST models, the negative and positive classes for each outcome were also randomly split into five-fold 80:20 train:validation splits and individually used across five

model trial runs. Model results for internal validation are summarized in Table 4.2 for SSI and CST with the relevant metrics of average training accuracy, training loss, validation accuracy, validation loss, and ROC across each fold as well as the average score for the best epoch prior to early stopping. From the SSI model results, overall model training accuracy (89.310%) and validation accuracy (80.833%) with similarly low loss functions (0.255 and 0.222, respectively) demonstrated promising performance on generalizing within the Atrium sample. These results are further supported with high calculated ROC in the validation sample averaging 0.898 across five trial runs, indicating the model has high internal discriminative ability.

In the CST model, the training accuracy reflected the classification accuracy from LOOCV, however, validation accuracy in this sample was notably lower at 66.552% (validation loss: 0.699) compared to LOOCV. In addition, validation ROC was very poor with a reported value of 0.509 across all five trial runs indicating very poor discriminative ability on the internal sample.

Table 4.2: Best training and validation accuracy by fold as well as the overall average across folds for CST and SSI outcomes

Run	**1**	**2**	**3**	**4**	**5**	**Average**
SSI Outcome						
Epoch	89	96	99	88	97	93.8
Training Acc.	87.586%	90.690%	91.724%	89.310%	87.241%	89.310%
Training Loss	0.288	0.241	0.217	0.278	0.254	0.255
Val. Acc.	81.945%	81.942%	79.166%	79.167%	81.945%	80.833%
Val. Loss	0.221	0.218	0.243	0.241	0.190	0.222
Val. ROC	0.966	0.899	0.897	0.865	0.863	0.898
CST Outcome						
Epoch	93	96	96	97	97	95.8
Training Acc.	73.2218%	74.4770%	75.7322%	76.1506%	76.9854%	75.3134%
Training Loss	0.4713	0.4237	0.4307	0.4122	0.4166	0.4309
Val. Acc.	60.3448%	70.6897%	65.5172%	67.2414%	68.9655	66.5517%
Val. Loss	0.7126	0.6671	0.7368	0.6886	0.6906	0.6991
Val. ROC	0.4388	0.5667	0.4597	0.5320	0.5486	0.5092

External Validation

Test validation consisted of generating 10 random subsamples (n= 50) from the 75-sample external patient set and calculating the corresponding accuracy for each sample. In addition, full test sample results for accuracy and ROC with corresponding graph are displayed in table 4.3 below. For the SSI model, the classification accuracy had relatively good performance in both randomly selected 50-patient subsamples as well as the total sample.

In CST, external validation accuracy was similar across subsample patient cohorts (63.8%) and the full 75-person cohort (65.33%). Furthermore, these results corroborate the observed internal validation accuracy from the atrium set (66.552% average).

Table 4.3: Accuracy measures from 10 image subsamples (n=50 per subsample) and overall (n=75) average accuracy and ROC in CST and SSI outcomes

	Subsample Average	**Overall Average**
SSI Outcome		
Accuracy	72.80%	73.33%
CST Outcome		
Accuracy	63.80%	65.33%

Pooled cross-validation Using Internal and External Samples

Since the initial external validation for both CST and SSI models reported poor discriminative ability measured by ROC. To ensure that the results obtained from this pooled validation were not solely due to the larger sample size, each patient sample used to build the CST and SSI model respectively were proportionately sampled to have the same total population as the original internal sample. In other words, the total sample comprising Atrium and Ohio patients for the SSI model would represent 362 patients and 297 for the CST model. For SSI, the sample consisted of 300 (236 negative, 64 positive) and 62 Ohio (51 negative, 11 positive)

patients for internal validation. The CST model therefore consisted of 237 (160 negative, 77 positive) Atrium patients and 60 (22 negative, 38 positive) Ohio patients for internal validation. Similar to the internal validation conducted above, these samples were split into 80:20 train: test samples and individually ran across five model trial runs. Model performance results by trial run from the CST model internal training with both Atrium and Ohio data are summarized in table 4.4 below.

Based on the SSI model pooled validation results, training accuracy was very high averaging 97.923% across five trial runs (training loss: 0.058). Furthermore, the validation accuracy was much higher in the SSI model averaging 88.611% (validation loss: 0.414) indicating much better model generalizability. Average model ROC value also reflects this result with a reported average score of 0.876, suggesting strong model discriminative ability and generalizability when coupled with the reported high calculated accuracy.

In contrast, the CST validation results continues to be poor with a less than 40% validation across averaged across the five trial runs (validation loss: 1.104), suggesting very low generalizability. Average model training accuracy was higher at 91.261% (training loss: 0.097). However, coupled with the low validation accuracy, this indicates that the model is likely overfitting on the training data. In addition, validation ROC supports this low accuracy score with an average value of 0.568, indicating minimal model discriminative ability.

Table 4.4: Best training and validation accuracy by trial run as well as the overall average across folds for CST and SSI using all samples (Atrium and Ohio)

Run	**1**	**2**	**3**	**4**	**5**	**Average**
SSI Outcome						
Epoch	79	86	94	69	82	82
Training Acc.	97.924%	97.924%	98.267%	97.924%	97.578%	97.923%
Training Loss	0.065	0.047	0.049	0.069	0.060	0.058
Val. Acc.	86.111%	87.500%	91.667%	87.500%	90.278%	88.611%

Val. Loss	0.432	0.432	0.454	0.386	0.365	0.413
Val. ROC	0.880	0.862	0.853	0.881	0.903	0.876
CST Outcome						
Epoch	79	90	86	95	80	86
Training Acc.	92.017%	90.336%	91.177%	91.176%	91.597%	91.261%
Training Loss	0.095	0.109	0.103	0.097	0.084	0.097
Val. Acc.	37.647%	42.353%	41.177%	38.824	37.647%	39.529%
Val. Loss	1.158	0.988	1.124	1.220	1.032	1.104
Val. ROC	0.572	0.579	0.534	0.540	0.614	0.568

4.4 Discussion

Our study aimed to externally validate a pretrained model architecture built on an AWR patient cohort from a tertiary hernia center in the South Piedmont area of the United States (Elhage et. al., 2021). In this endeavor, we internally trained and validated using the ResNet-18 model architecture for outcomes: CST and SSI. From the validation, initial results for both outcome models from leave-one-out cross validation showed promise with high proportions of the validation data correctly classified. Furthermore, in the internal cross validation, the SSI model had high validation accuracy with greater than 80% of unseen cases correctly classified.

However, the discriminative ability of both models after internal validation was found to be relatively poor with ROC values of approximately 0.6 for the CST model suggesting poor generalization in predictive performance from the initial models. This performance discrepancy may be due to surgical divisions across the United States having highly varied patient populations in terms of severity of patients' conditions, types and complexity of performed procedures, and established protocols and standards of care (Poulose et. al., 2016). In addition, surgical decision-making in utilizing the CST varies significantly with many incorporating different criteria and modifications to the procedure (Adekunle et. al., 2013). Such variability among surgery practices may also contribute to the disparity in model discriminative ability between the Atrium and Ohio cohorts employed in the study. As a result, a future investigation

may warrant supplementing patient image data with other tabular patient data, i.e patient demographic characteristics, lab values, etc. in an effort to build more promising classification models.

As a result, we investigated pooling the external sample data together with the initial internal sample data to perform one large validation study with each cohort's data available to the model during training. Equal proportions of internal and external sample data were set aside for cross validation and left unseen to the model. Similar to the initial research investigation, the CST model demonstrated poor performance in the validation set with low validation accuracy at approximately 40% as well as poor ROC value at 0.6. However, the SSI model reported much higher validation accuracy with greater than 85% of cases in the validation sample correctly classified and a calculated ROC value of 0.88. Although unsuccessful with the CST model, the results from the SSI model demonstrate that there is some possibility in creating externally generalizable models for prediction in AWR surgery patients.

Other classification models in hernia surgery patients have achieved internal validation accuracies of approximately 70% for identifying if a patient will develop SSI after surgery representing the general standard for prediction models in hernia surgery (Hassan et. al., 2022). Our developed model represents a significant improvement on existing classification models for determining whether a patient will develop an SSI event. With SSI events costing approximately $11,000 compounded by the reduced quality of life and toll on physical health, minimizing these possible events by identifying those most at risk prior to surgery can assist physicians in optimizing preoperative and postoperative treatment plans for these high-risk patients (Wilson & Farooque, 2022).

Correctly identifying patients who require CST procedures can reduce hernia recurrence rates by up to 20% (Luijendijk et. al., 2000). With little to no tools available to predict the probability of requiring a CST, creating validated models which can classify patients as requiring a CST has significant implications in improving patient physical health due to avoiding adverse health outcomes like the aforementioned hernia recurrence. In addition to the physical health costs, these procedures also come with financial costs averaging approximately $21,000 per patient in the United States (Davila et. al., 2016).

Some study limitations include the available sample data for both internal and external cohorts. Of note, proportions of classes between the internal and external cohorts were significantly skewed with approximately 33% of patients in the internal CST sample labeled as "positive" while 67% of cases in the external CST sample were labeled as "positive". This significant discrepancy in outcome proportions may have contributed to the poor predictive performance reported when measured on the external validation sample. A preliminary k-nearest neighbors cluster analysis was performed using the Ohio and Atrium patient images to determine if the two patient groups have defining features within a unique culture (Zhu et. al., 2022). These results are summarized by cluster in Appendix D. Future research could expand on the initially performed cluster analysis to potentially match patients with similar outcome severity in the internal and external cohorts. With limited data available for training and validation, utilizing more similarly matched cohorts could assist in building accurate and generalizable models to specific patient cohorts in surgery.

Another possible limitation is that this study was not an extensive search or analysis on possible deep learning model architectures and model fitting. There is the possibility that other available model architectures may better fit and classify the input data than ResNet-18 and, in

turn, have higher generalizability on external patient cohorts. Future research could include investigating multiple deep learning architectures to determine better performance between models on both internal and external sample data.

Our research contributions strengthen the available literature on externally validating existing deep learning prediction models in surgery. From the literature, other institutions have successfully validated deep learning-based prediction models in clinical research using external cohorts. A study by Lee et. al., (2020) demonstrated that multiple deep learning models utilized in cervical lymph node cancer diagnosis could be successfully validated and retain reasonable image accuracy (ranging from 0.784-0.884). Another study by Choi et. al., (2017) demonstrated similar validation accuracy, this time predicting thyroid cancer nodules with AUC values reported at 0.83. These promising results demonstrate that deep learning models can successfully generalize to other clinical populations beyond those utilized in the training sample without significant drops in classification accuracy due to limitations from overfitting or other potential model shortfalls.

In conclusion, we successfully trained and validated deep learning models for patients undergoing abdominal wall reconstruction surgery. Although the initially assessed external performance was poor, the pooled validation results and cluster analysis show promise in successfully building a generalizable deep learning model for predicting patient outcomes in surgery. Supplementing this and other studies with external validation is imperative in determining model viability for implementation in the clinical setting to aid clinical decision-making. Model integration into clinical practice can therefore assist surgeons in determining high-risk patients for surgical complexity or postoperative outcomes and provide a more

objective means of tiering patient risk. Those identified can consequently be managed through specialized interventions to improve patient outcomes, satisfaction, and quality of life.

CHAPTER 4 REFERENCE LIST

Adekunle, S., Pantelides, N. M., Hall, N. R., Praseedom, R., & Malata, C. M. (2013). Indications and outcomes of the components separation technique in the repair of complex abdominal wall hernias: experience from the cambridge plastic surgery department. Eplasty, 13, e47–e47.

Bahl, M. (2020). Artificial Intelligence: A Primer for Breast Imaging Radiologists. *Journal of Breast Imaging* (Online), 2(4), 304–314. https://doi.org/10.1093/jbi/wbaa033

Chen, H.-L., Yu, S.-J., Xu, Y., Yu, S.-Q., Zhang, J.-Q., Zhao, J.-Y., Liu, P., & Zhu, B. (2018). Artificial Neural Network: A Method for Prediction of Surgery-Related Pressure Injury in Cardiovascular Surgical Patients. *Journal of Wound, Ostomy, and Continence Nursing*, 45(1), 26–30. https://doi.org/10.1097/WON.0000000000000388

Choi, Y. J., Baek, J. H., Park, H. S., Shim, W. H., Kim, T. Y., Shong, Y. K., & Lee, J. H. (2017). A Computer-Aided Diagnosis System Using Artificial Intelligence for the Diagnosis and Characterization of Thyroid Nodules on Ultrasound: Initial Clinical Assessment. *Thyroid* (New York, N.Y.), 27(4), 546–552. https://doi.org/10.1089/thy.2016.0372

Collins, G. S., Reitsma, J. B., Altman, D. G., & Moons, K. G. M. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ* (Online), 350(jan07 4), g7594–g7594. https://doi.org/10.1136/bmj.g7594

Davila, D. G., Parikh, N., Frelich, M. J., & Goldblatt, M. I. (2016). The increased cost of ventral hernia recurrence: a cost analysis. *Hernia: the Journal of Hernias and Abdominal Wall Surgery*, 20(6), 811–817. https://doi.org/10.1007/s10029-016-1515-5

Elhage, S. A., Deerenberg, E. B., Ayuso, S. A., Murphy, K. J., Shao, J. M., Kercher, K. W., Smart, N. J., Fischer, J. P., Augenstein, V. A., Colavita, P. D., & Heniford, B. T. (2021). Development and Validation of Image-Based Deep Learning Models to Predict Surgical Complexity and Complications in Abdominal Wall Reconstruction. *Archives of Surgery* (Chicago. 1960), 156(10), 933–. https://doi.org/10.1001/jamasurg.2021.3012

Fatemi, P., Zhang, Y., Han, S. S., Purington, N., Zygourakis, C. C., Veeravagu, A., Desai, A., Park, J., Shuer, L. M., & Ratliff, J. K. (2022). External validation of a predictive model of adverse events following spine surgery. *The Spine Journal*, 22(1), 104–112. https://doi.org/10.1016/j.spinee.2021.06.006

Geron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. O'Reilly Media, Incorporated.

Hassan, A. M., Lu, S.-C., Asaad, M., Liu, J., Offodile, A. C. I., Sidey-Gibbons, C., & Butler, C. E. (2022). Novel Machine Learning Approach for the Prediction of Hernia Recurrence, Surgical Complication, and 30-Day Readmission after Abdominal Wall Reconstruction. *Journal of the American College of Surgeons*, 234(5), 918. https://doi.org/10.1097/XCS.0000000000000141

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied Logistic Regression (3. Aufl.). Wiley.

Iiduka, H. (2022). Appropriate Learning Rates of Adaptive Learning Rate Optimization Algorithms for Training Deep Neural Networks. *IEEE Transactions on Cybernetics*, 52(12), 1–12. https://doi.org/10.1109/TCYB.2021.3107415

Lee, J. H., Ha, E. J., Kim, D., Jung, Y. J., Heo, S., Jang, Y., An, S. H., & Lee, K. (2020). Application of deep learning to the diagnosis of cervical lymph node metastasis from thyroid cancer with CT: external validation and clinical utility for resident training. *European Radiology*, 30(6), 3066–3072. https://doi.org/10.1007/s00330-019-06652-4

Luijendijk, R. W., Hop, W. C. ., van den Tol, M. P., de Lange, D. C. ., Braaksma, M. M. ., IJzermans, J. N. ., … Jeekel, J. (2000). A Comparison of Suture Repair with Mesh Repair for Incisional Hernia. *The New England Journal of Medicine*, 343(6), 392–398. https://doi.org/10.1056/NEJM200008103430603

Meijer, R. J., & Goeman, J. J. (2013). Efficient approximate k-fold and leave-one-out cross-validation for ridge regression. *Biometrical Journal*, 55(2), 141–155. https://doi.org/10.1002/bimj.201200088

Montagnon, E., Cerny, M., Cadrin-Chênevert, A., Hamilton, V., Derennes, T., Ilinca, A., Vandenbroucke-Menu, F., Turcotte, S., Kadoury, S., & Tang, A. (2020). Deep learning workflow in radiology: a primer. *Insights into Imaging*, 11(1), 22–22. https://doi.org/10.1186/s13244-019-0832-5

Navarro, F., Dapper, H., Asadpour, R., Knebel, C., Spraker, M. B., Schwarze, V., Schaub, S. K., Mayr, N. A., Specht, K., Woouff, H. C., Lambin, P., Gersing, A. S., Nyflot, M. J., Menze, B. H., Combs, S. E., & Peeken, J. C. (2021). Development and External Validation of Deep-Learning-Based Tumor Grading Models in Soft-Tissue Sarcoma Patients Using MR Imaging. *Cancers*, 13(12), 2866–. https://doi.org/10.3390/cancers13122866

Ramspek, C. L., Jager, K. J., Dekker, F. W., Zoccali, C., & van Diepen, M. (2021). External validation of prognostic models: What, why, how, when and where? *Clinical Kidney Journal*, 14(1), 49–58. https://doi.org/10.1093/ckj/sfaa188

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., … Chintala, S. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. https://doi.org/10.48550/arxiv.1912.01703

Savadjiev, P., Chong, J., Dohan, A., Vakalopoulou, M., Reinhold, C., Paragios, N., & Gallix, B. (2019). Demystification of AI-driven medical image interpretation: past, present and future. *European Radiology*, 29(3), 1616–1624. https://doi.org/10.1007/s00330-018-5674-x

Siontis, G. C. ., Tzoulaki, I., Castaldi, P. J., & Ioannidis, J. P. . (2015). External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. *Journal of Clinical Epidemiology*, 68(1), 25–34. https://doi.org/10.1016/j.jclinepi.2014.09.007

Starke, S., Leger, S., Zwanenburg, A., Leger, K., Lohaus, F., Linge, A., Schreiber, A., Kalinauskaite, G., Tinhofer, I., Guberina, N., Guberina, M., Balermpas, P., von der Grun, J., Ganswindt, U., Belka, C., Peeken, J. C., Combs, S. E., Boeke, S., Zips, D., … Loeck, S. (2020). 2D and 3D convolutional neural networks for outcome modelling of locally advanced head and neck squamous cell carcinoma. *Scientific Reports*, 10(1), 15625–15625. https://doi.org/10.1038/s41598-020-70542-9

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2013). Intriguing properties of neural networks. https://doi.org/10.48550/arxiv.1312.6199

Wilson, R. B., & Farooque, Y. (2022). Risks and Prevention of Surgical Site Infection After Hernia Mesh Repair and the Predictive Utility of ACS-NSQIP. *Journal of Gastrointestinal Surgery*, 26(4), 950–964. https://doi.org/10.1007/s11605-022-05248-6

Yamamoto, T., Takahashi, S., Ichihara, K., Hiyama, Y., Uehara, T., Hashimoto, J., Hirobe, M., & Masumori, N. (2015). How do we understand the disagreement in the frequency of surgical site infection between the CDC and Clavien-Dindo classifications? *Journal of Infection and Chemotherapy: Official Journal of the Japan Society of Chemotherapy*, 21(2), 130–133. https://doi.org/10.1016/j.jiac.2014.10.016

Zhu, H., Wang, X., & Wang, R. (2022). Fuzzy Monotonic K-Nearest Neighbor Versus Monotonic Fuzzy K-Nearest Neighbor. *IEEE Transactions on Fuzzy Systems*, 30(9), 3501–3513. https://doi.org/10.1109/TFUZZ.2021.3117450

APPENDIX D: K-NEAREST NEIGHBORS CLUSTER ANALYSIS APPROACH ON PATIENT CT IMAGES FOR INTERNAL AND EXTERNAL COHORTS

Wound model:
Clusters = 2

Cluster #	**# of images**
Ohio Dataset	
1	6
2	69
Atrium Dataset	
1	178
2	184

Clusters = 3

Cluster #	**# of images**
Ohio Dataset	
1	8
2	63
3	4
Atrium Dataset	
1	149
2	141
3	72

Clusters = 4

Cluster #	**# of images**
Ohio Dataset	
1	1
2	29
3	41
4	3
Atrium Dataset	
1	73
2	135
3	65
4	89

Component Separation model:
Clusters = 2

Cluster #	**# of images**
Ohio Dataset	
1	5

2	70
Atrium Dataset	
1	181
2	116

Clusters = 3

Cluster #	**# of images**
Ohio Dataset	
1	5
2	66
3	4
Atrium Dataset	
1	121
2	85
3	91

Clusters = 4
Ohio Dataset

Cluster #	**# of images**
Ohio Dataset	
1	35
2	1
3	36
4	3
Atrium Dataset	
1	102
2	89
3	41
4	65

CHAPTER 5: CONCLUSIONS

This dissertation investigated multiple statistical and machine learning techniques within the context of MIS. In Chapter 2, we investigated patient preclinical variables to determine what factors were significantly associated with patient QOL six months after surgery. From the study, 20 unique preclinical variables were identified with a significant association to patient QOL after surgery. Of these variables four were noted as risk factors of statistical importance with patient QOL including patient diabetes status, patient history of cholecystectomy, patient history of appendectomy, and patient preoperative CCS score. These findings were further validated by stratified k-fold cross validation. This first study demonstrated the utility of accurate predictive models in aiding clinicians during preoperative assessments by determining those most at risk for reduced QOL after surgery.

Chapter 3 investigates image data modeled using complex deep learning architectures to predict patient QOL after surgery. Given patient images prior to surgery are relatively limited, maximizing the available data represents the study's primary focus. Multiple qualitative methodologies were investigated to improve image quality. By optimizing available image data, researchers can maximize the available data to build predictive models that otherwise would not have been possible. In addition, by reducing the noise and variation in images, researchers can provide more meaningful and accurate predictions based on the input data. Ultimately, this study demonstrated that cropping, aligning, and creating image composites from the available image data on the patient produced the most optimal input data to predict patient QOL.

Finally, Chapter 4 sought to validate an existing deep learning model in MIS designed to classify two postoperative outcomes following AWR surgery. Research findings suggest that although the generated predictive model externally classified probability of SSI with reasonable

accuracy, efforts to create a validated model to predict the probability of CST were unsuccessful. External validation is imperative in determining model viability for implementation in the clinical setting; otherwise, model predictions cannot be trusted as a supplement in the clinical decision-making process.

In Chapter 4, case-matching on internal and external patient cohorts on features including surgical complexity or procedure severity warrants further exploration. Surgical divisions across the United States may have highly varied patient populations in terms of severity of patients' conditions, types and complexity of performed procedures, and established protocols and standards of care (Poulose et. al., 2016). Effectively validating model generalizability on other cohorts is exceedingly difficult due to the limitations in training and validation sample sizes (Krammer et. al., 2022). By matching based on these aforementioned criteria, researchers may have greater insight into model performance on external patient groups, controlling for some inherent variability present within clinical practices.

Future studies would benefit from investigating additional deep learning architectures to effectively model and classify other outcomes in AWR surgery. With a multitude of available architectures and updated versions constantly under development, other models may more effectively fit patient CT data for outcome classification (Khamparia & Singh, 2019). In addition, other clinical studies have had significant success in creating ensemble models utilizing both patient row-level data (for example, patient demographic characteristics) and image data to classify outcomes which could be applicable within this study context (Suk et. al., 2017).

Taken together, these applied statistics, machine learning, and deep learning techniques represent an effective means to optimize patient treatment plans and QOL before and after MIS. However, due to the inherent complexity and nuance surrounding patient image and tabular data,

these methodologies must be appropriately modified to fit the clinical context. Shortcomings such as missing data and image noise and variation pose a substantial barrier to clinical researchers seeking to build and validate predictive models in medicine. Recognizing and controlling these context-specific issues grants clinical researchers the ability to optimize the construction of robust predictive models that are both accurate and generalizable to multiple patient cohorts. Clinicians and clinical researchers can combine their respective domain expertise into actionable and valid products to provide treatments designed to optimize patient-specific care. In conclusion, to effectively maximize applied statistics, machine learning, and deep learning methods in surgery, clinical context-specific knowledge is critical in their respective application and eventual implementation.

GENERAL REFERENCES

Akkus, Z., Galimzianova, A., Hoogi, A., Rubin, D. L., & Erickson, B. J. (2017). Deep Learning for Brain MRI Segmentation: State of the Art and Future Directions. *Journal of Digital Imaging*, 30(4), 449–459. https://doi.org/10.1007/s10278-017-9983-4

Balaji, T., & Sumathi, M. (2014). Effective Features of Remote Sensing Image Classification Using Interactive Adaptive Thresholding Method. arXiv.org.

Bekelis, K., Desai, A., Bakhoum, S. F., & Missios, S. (2014). A predictive model of complications after spine surgery: the National Surgical Quality Improvement Program (NSQIP) 2005–2010. *The Spine Journal*, 14(7), 1247–1255. https://doi.org/10.1016/j.spinee.2013.08.009

Bendifallah, S., Daraï, E., & Ballester, M. (2016). Predictive Modeling: A New Paradigm for Managing Endometrial Cancer. *Annals of Surgical Oncology*, 23(3), 975–988. https://doi.org/10.1245/s10434-015-4924-2

Chen, H.-N., & Tsai, Y.-F. (2013). A predictive model for disability in patients with lumbar disc herniation. *Journal of Orthopaedic Science: Official Journal of the Japanese Orthopaedic Association*, 18(2), 220–229. https://doi.org/10.1007/s00776-012-0354-1

Chollet, F., Deep Learning with Python. Shelter Island, NY, USA: Manning, 2017.

Tian, C., Xu, Y., & Zuo, W. (2020). Image denoising using deep CNN with batch renormalization. *Neural Networks*, 121, 461–473. https://doi.org/10.1016/j.neunet.2019.08.022

Elhage, S. A., Deerenberg, E. B., Ayuso, S. A., Murphy, K. J., Shao, J. M., Kercher, K. W., Smart, N. J., Fischer, J. P., Augenstein, V. A., Colavita, P. D., & Heniford, B. T. (2021). Development and Validation of Image-Based Deep Learning Models to Predict Surgical Complexity and Complications in Abdominal Wall Reconstruction. *Archives of Surgery* (Chicago. 1960), 156(10), 933–. https://doi.org/10.1001/jamasurg.2021.3012

Fitzpatrick, R., Davey, C., Buxton, M. J., & Jones, D. R. (1998). Evaluating patient-based outcome measures for use in clinical trials. *Health technology assessment* (Winchester, England), 2(14), i–74.

Heniford, B. T., Walters, A. L., Lincourt, A. E., Novitsky, Y. W., Hope, W. W., & Kercher, K. W. (2008). Comparison of generic versus specific quality-of-life scales for mesh hernia repairs. *Journal of the American College of Surgeons*, 206(4), 638–644.

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied Logistic Regression (3. Aufl.). Wiley.

Jia Deng, Wei Dong, Socher, R., Li-Jia Li, Kai Li, & Li Fei-Fei. (2009). ImageNet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255. https://doi.org/10.1109/CVPR.2009.5206848

Khamparia, A., & Singh, K. M. (2019). A systematic review on deep learning architectures and applications. *Expert Systems*, 36(3), e12400–n/a. https://doi.org/10.1111/exsy.12400

Krammer, S., Li, Y., Jakob, N., Boehm, A. S., Wolff, H., Tang, P., Lasser, T., French, L. E., & Hartmann, D. (2022). Deep learning-based classification of dermatological lesions given a limited amount of labelled data. *Journal of the European Academy of Dermatology and Venereology*, 36(12), 2516–2524. https://doi.org/10.1111/jdv.18460

Lee, Y. H., Bang, H., & Kim, D. J. (2016). How to Establish Clinical Prediction Models. *Endocrinology and metabolism* (Seoul, Korea), 31(1), 38–44. https://doi.org/10.3803/EnM.2016.31.1.38

Lo, J., Cardinell, J., Costanzo, A., & Sussman, D. (2021). Medical Augmentation (Med-Aug) for Optimal Data Augmentation in Medical Deep Learning Networks. *Sensors* (Basel, Switzerland), 21(21), 7018–. https://doi.org/10.3390/s21217018

Luijendijk, R. W., Hop, W. C. ., van den Tol, M. P., de Lange, D. C. ., Braaksma, M. M. ., IJzermans, J. N. ., Boelhouwer, R. U., de Vries, B. C., Salu, M. K. ., Wereldsma, J. C. ., Bruijninckx, C. M. ., & Jeekel, J. (2000). A Comparison of Suture Repair with Mesh Repair for Incisional Hernia. *The New England Journal of Medicine*, 343(6), 392–398. https://doi.org/10.1056/NEJM200008103430603

Montes Pérez, A., Roca, G., Sabaté, S., Lao, J. I., Navarro i Cuartiellas, A., Cantillo, J., Canet, J., & GENDOLCAT Study Group. (2015). Genetic and Clinical Factors Associated with Chronic Postsurgical Pain after Hernia Repair, Hysterectomy, and Thoracotomy: A Two-year Multicenter Cohort Study. https://doi.org/10.1097/ALN.0000000000000611

Osorio, J. A., Scheer, J. K., & Ames, C. P. (2016). Predictive modeling of complications. *Current reviews in musculoskeletal medicine*, 9(3), 333–337. https://doi.org/10.1007/s12178-016-9354-7

Poulose, B. K., Shelton, J., Phillips, S., Moore, D., Nealon, W., Penson, D., Beck, W., & Holzman, M. D. (2012). Epidemiology and cost of ventral hernia repair: making the case for hernia research. *Hernia: the Journal of Hernias and Abdominal Wall Surgery*, 16(2), 179–183. https://doi.org/10.1007/s10029-011-0879-9

Poulose, B. K., Roll, S., Murphy, J. W., Matthews, B. D., Todd Heniford, B., Voeller, G., Hope, W. W., Goldblatt, M. I., Adrales, G. L., & Rosen, M. J. (2016). Design and implementation of the Americas Hernia Society Quality Collaborative (AHSQC): improving

value in hernia care. *Hernia: the Journal of Hernias and Abdominal Wall Surgery*, 20(2), 177–189. https://doi.org/10.1007/s10029-016-1477-7

Ramspek, C. L., Jager, K. J., Dekker, F. W., Zoccali, C., & van Diepen, M. (2021). External validation of prognostic models: What, why, how, when and where? *Clinical Kidney Journal*, 14(1), 49–58. https://doi.org/10.1093/ckj/sfaa188

Rivas-Blanco, I., Perez-Del-Pulgar, C. J., Garcia-Morales, I., & Munoz, V. F. (2021). A Review on Deep Learning in Minimally Invasive Surgery. *IEEE Access*, 9, 48658–48678. https://doi.org/10.1109/ACCESS.2021.3068852

Schug, S. A., & Bruce, J. (2017). Risk stratification for the development of chronic postsurgical pain. *Pain Reports*, 2(6), e627–e627. https://doi.org/10.1097/PR9.0000000000000627

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on Image Data Augmentation for Deep Learning. Journal of Big Data, 6(1), 1–48. https://doi.org/10.1186/s40537-019-0197-0

Shrestha, A., & Mahmood, A. (2019). Review of Deep Learning Algorithms and Architectures. *IEEE Acces*s, 7, 1–1. https://doi.org/10.1109/ACCESS.2019.2912200

Simonyan, K., & Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv.org.

Suk, H.-I., Lee, S.-W., & Shen, D. (2017). Deep ensemble learning of sparse regression models for brain disease diagnosis. *Medical Image Analysis*, 37, 101–113. https://doi.org/10.1016/j.media.2017.01.008

Tang, C. L., & Schlich, T. (2017). Surgical Innovation and the Multiple Meanings of Randomized Controlled Trials: The First RCT on Minimally Invasive Cholecystectomy (1980–

2000). *Journal of the History of Medicine and Allied Sciences*, 72(2), 117–141. https://doi.org/10.1093/jhmas/jrw027

Urbach, D. R. (2005). Measuring Quality of Life After Surgery. Surgical Innovation, 12(2), 161–165. https://doi.org/10.1177/155335060501200216

Velanovich, V. (1998). Comparison of generic (SF-36) vs. disease-specific (GERD-HRQL) quality-of-life scales for gastroesophageal reflux disease. *Journal of Gastrointestinal Surgery*, 2(2), 141–145. https://doi.org/10.1016/S1091-255X(98)80004-8

Ware, J., & Sherbourne, C. (1992). The MOS 36-Item Short-Form Health Survey (SF-36) .1. Conceptual-Framework and Item Selection. *Medical Care*, 30(6), 473–483. https://doi.org/10.1097/00005650-199206000-00002

Wang, F., Casalino, L. P., & Khullar, D. (2019). Deep Learning in Medicine—Promise, Progress, and Challenges. *JAMA Internal Medicine*, 179(3), 293–294. https://doi.org/10.1001/jamainternmed.2018.7117

Wang, J., Yang, X., Cai, H., Tan, W., Jin, C., & Li, L. (2016). Discrimination of Breast Cancer with Microcalcifications on Mammography by Deep Learning. *Scientific Reports*, 6(1), 27327–27327. https://doi.org/10.1038/srep27327

Wechter, M., Pearlman, M., & Hartmann, K. (2005). Reclosure of the disrupted laparotomy wound - A systematic review. *Obstetrics and Gynecology* (New York. 1953), 106(2), 376–383. https://doi.org/10.1097/01.AOG.0000171114.75338.06

Xiao, M., Wu, Y., Zuo, G., Fan, S., Yu, H., Shaikh, Z. A., & Wen, Z. (2021). Addressing Overfitting Problem in Deep Learning-Based Solutions for Next Generation Data-Driven Networks. Wireless Communications and Mobile Computing, 2021, 1–10. https://doi.org/10.1155/2021/8493795

Zhu, X.-L., Shen, H.-B., Sun, H., Duan, L.-X., & Xu, Y.-Y. (2022). Improving segmentation and classification of renal tumors in small sample 3D CT images using transfer learning with convolutional neural networks. *International Journal for Computer Assisted Radiology and Surgery*, 17(7), 1303–1311. https://doi.org/10.1007/s11548-022-02587-2