Organ Viability Assessment in Transplantation based on Data-driven Modeling

Qing Lan

Dissertation Submitted to the Faculty of the Virginia Polytechnic Institute and State University in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy In Industrial and Systems Engineering

Ran Jin, Chair
Nathan Lau
John L. Robertson
Christian Wernz

January 22, 2020
Blacksburg, VA

Keywords: Data-driven Decision-making, Non-invasive Assessment, Organ Transplantation, Statistical Modeling
Organ Viability Assessment in Transplantation based on Data-driven Modeling
Qing Lan

ABSTRACT
Organ transplantation is one of the most important and effective solutions to save end-stage patients, who have one or more critical organ failures. However, the inadequate organs for transplantation to meet the demands has been the major issue. Even worse, the lack of accurate non-invasive assessment methods wastes 20% of donor organs every year. Currently, the most frequently used organ assessment methods are visual inspections and biopsy. Yet both methods are subjective: the assessment accuracy depends on the evaluator’s experience. Moreover, repeating biopsies will potentially damage the organs. To reduce the waste of donor organs, non-invasive, online, and quantitative organ assessment methods are in great needs.

Organ viability assessment is a challenging issue due to four reasons: 1) there are no universally accepted guidelines or procedures for surgeons to quantitatively assess the organ viability; 2) there is no easy-deployed and non-invasive biological in situ data to correlate with organ viability; 3) the organs viability is difficult to model because of heterogeneity among organs; 4) both visual inspection and biopsy can be applied only at present time, and how to forecast the viability of similar-but-non-identical organs at a future time is still in shadow.

Motivated by the challenges, the overall objective of this dissertation is to develop non-invasive and quantitative online assessment methods to predict and forecast the organ viability. As a result, four data-driven modeling research tasks are investigated to achieve the overall objective:
1) Quantitative and qualitative models are used to jointly predict the number of dead cells and the liver viability based on features extracted from biopsy images. This method can quantitatively
assess the organ viability, which could be used to validate the biopsy results from pathologists to increase the evaluation accuracy.

2) A multitask learning logistic regression model is applied to assess liver viability by using principal component analysis to extract infrared image features to quantify the correlation between liver viability and spatial infrared imaging data. This non-invasive online assessment method can evaluate the organ viability without physical contact to reduce the risk of damaging the organs.

3) A spatial-temporal smooth variable selection method is conducted to improve the liver viability prediction accuracy by considering both spatial and temporal effects from the infrared images without feature engineering. In addition, it provides medical interpretation based on variable selection to highlight the most significant regions on the liver resulting in viability loss.

4) A multitask general path model is implemented to forecast the heterogeneous kidney viability based on limited historical data by learning the viability loss paths of each kidney during preservation. The generality of this method is validated by tissue deformation forecasting in needle biopsy process to potentially improve the biopsy accuracy.

In summary, the proposed data-driven methods can predict and forecast the organ viability without damaging the organ. As a result, the increased utilization rate of donor organs will benefit more end-stage patients by dramatically extending their life spans.
Organ Viability Assessment in Transplantation based on Data-driven Modeling
Qing Lan

GENERAL AUDIENCE ABSTRACT
Organ transplantation is the ultimate solution to save end-stage patients with one or more organ failures. However, the inadequate organs for transplantation to meet the demands has been the major issue. Even worse, the lack of accurate and non-invasive viability assessment methods wastes 20% of donor organs every year. Currently, the most frequently used organ assessment methods are visual inspections and biopsy. Yet both methods are subjective: the assessment accuracy depends on the personal experience of evaluator. Moreover, repeating biopsies will potentially damage the organs. As a result, online non-invasive and quantitative organ assessment methods are in great needs. It is extremely important because such methods will increase the organ utilization rate by saving more discarded organs with transplantation potential.

The overall objective of this dissertation is to advance the knowledge on modeling organ viability by developing online non-invasive and quantitative methods to predict and forecast the viability of heterogeneous organs in transplantation. After an introduction in Chapter 1, four research tasks are investigated. In Chapter 2, quantitative and qualitative models jointly predicting porcine liver viability are proposed based on features from biopsy images to validate the biopsy results. In Chapter 3, a multi-task learning logistic regression model is proposed to assess the cross-liver viability by correlating liver viability with spatial infrared data validated by porcine livers. In Chapter 4, a spatial-temporal smooth variable selection is proposed to predict liver viability by considering both spatial and temporal correlations in modeling without feature engineering, which is also validated by porcine livers. In addition, the variable selection results provide medical interpretations by capturing the significant regions on the liver in predicting viability. In Chapter
5, a multitask general path model is proposed to forecast kidney viability validated by porcine kidney. This forecasting method is generalized to apply to needle biopsy tissue deformation case study with the objective to improve the needle insertion accuracy. Finally, I summarize the research contribution and discuss future research directions in Chapter 6. The proposed data-driven methods can predict and forecast organ viability without damaging the organ. As a result, the increased utilization rate of donor organs will benefit more patients by dramatically extending their life spans and bringing them back to normal daily activities.
To my lost youth
ACKNOWLEDGEMENT

I would like to express my heartfelt gratitude to my Ph.D. advisor Professor Ran Jin, for providing me this great opportunity to involve in several research projects to explore unknowns in his lab. My Ph.D. journey would not have been possible without the support, encouragement, and supervision from him. I am very grateful to my other committee members, Professor Nathan Lau, Professor John Robertson, and Professor Christian Wernz for their insightful suggestions, experiment guidance, supervision and support during my Ph.D. study.

In addition, I would like to thank Professor Xinwei Deng and Professor Yong Lei for advising in collaborative research paper writing, and experimental planning of the needle biopsy process, respectively.

I would like to thank my lab members and friends for the supports and encouragements during my Ph.D. study, including but not limited to: Mr. Xiaoyu Chen, Mr. Jooneun Choi, Dr. SungKu Kang, Mr. Abhishek Kar, Ms. Jingran Li, Ms. Murong Li, Mr. Yifu Li, Dr. Chen’ang Liu, Dr. Jia Liu, Ms. Karuniya Mohan, Dr. Gongzhuang Peng, Ms. Parshin Shojaee, Dr. Wenmeng Tian, Dr. Hongyue Sun, Ms. Yuxin Tong, Mr. Lening Wang, Dr. Wenjing Wang, Ms. Yingyan Zeng, Dr. Jie Zhang and Ms. Yutong Zhang, for their help academically and philosophically.

My deepest thanks and love extend to my grandparents, parents, parents-in-law, and sister for guiding me as a person, grandson, son, and brother to be a good man. My grandfather has left me during my study, but his spirit is always with me. Finally yet importantly, I wish to thank my dear wife, Dr. Yubin Shen, for extremely supporting my dream to pursue the Ph.D. and sharing every minute with me throughout this bittersweet journey. I love you so much.
# Table of Contents

Chapter 1. Introduction .................................................................................................................. 1  
  1.1. Motivation .......................................................................................................................... 1  
  1.2. State-of-the-Art ............................................................................................................... 2  
  1.3. Research Objectives ........................................................................................................ 4  
  1.4. Organization of the Dissertation ....................................................................................... 6  

Chapter 2. Quantitative and Qualitative Evaluation for Organ Preservation in Transplant ...... 7  
  2.1. Introduction ....................................................................................................................... 7  
  2.2. Research Background ...................................................................................................... 9  
  2.3. Proposed Methodology .................................................................................................... 11  
    2.3.1. Biopsy Procedure and Data Processing ...................................................................... 11  
    2.3.2. Quantitative and Qualitative Models .......................................................................... 13  
  2.4. Case Study ........................................................................................................................ 15  
    2.4.1. Prediction Performance .............................................................................................. 15  
    2.4.2. Variable Selection ...................................................................................................... 16  
  2.5. Conclusion and Future Work ........................................................................................... 18  

Chapter 3. Non-invasive Assessment of Liver Quality in Transplantation based on Thermal  
  Imaging Analysis ...................................................................................................................... 20  
  3.1. Introduction ....................................................................................................................... 20  
  3.2. Research Background ...................................................................................................... 22  
  3.3. Materials and Methods .................................................................................................... 24  
    3.3.1. Experimental Planning .............................................................................................. 24  
    3.3.2. Data Cleaning .......................................................................................................... 25  
    3.3.3. Proposed Method ..................................................................................................... 26  
  3.4. Results of Case Study ...................................................................................................... 31  
    3.4.1. Single Liver Evaluation ............................................................................................. 31  
    3.4.2. Cross-Liver Evaluation ............................................................................................. 32  
    3.4.3. PC Loading Analysis ................................................................................................. 37  
  3.5. Discussions ....................................................................................................................... 39  
  3.6. Conclusion ......................................................................................................................... 41  

Chapter 4. Modeling of Pre-transplant Liver Viability with Spatial-temporal Smooth Variable  
  Selection ...................................................................................................................................... 42
4.1. Introduction .................................................................................................................. 42
4.2. Background and Methods ............................................................................................ 43
   4.2.1. Research Background .......................................................................................... 43
   4.2.2. Experiment Design .............................................................................................. 46
   4.2.3. Dataset .............................................................................................................. 48
   4.2.4. The Proposed Method ........................................................................................ 49
4.3. Results and Discussions .............................................................................................. 53
4.4. Conclusions ................................................................................................................ 58

Chapter 5. Kidney Viability Forecasting by Multitask General Path Model in Transplantation . 60
5.1. Introduction ................................................................................................................ 60
5.2. Research Background ............................................................................................... 62
5.3. Materials and Methods ............................................................................................. 65
   5.3.1. The Proposed Method ....................................................................................... 65
   5.3.2. Case Study 1 ..................................................................................................... 68
   5.3.3. Case Study 2 ..................................................................................................... 71
5.4. Results and Discussions ............................................................................................ 72
5.5. Conclusions ................................................................................................................ 80

Chapter 6. Conclusions and Future Research ..................................................................... 81

References ........................................................................................................................ 85
Appendixes ......................................................................................................................... 92
   Appendix A. Liver Quality Evaluation Results based on The Full Data Set (1431 Observations) ........................................................................................................................................ 92
   Appendix C. PC Loading Graphs for Livers #2, #3, and #4 ............................................ 102
   Appendix D. Image Pooling Results for Liver #2, #3, and #4 ........................................ 105
   Appendix E. Numerical Prediction Performance ............................................................ 106
   Appendix F. Liver #1 STSVS Variable Selection Results .................................................. 107
   Appendix G. Model Diagnostic and Discussions ............................................................. 108
   Appendix H. Future Research Discussions ...................................................................... 115
List of Figures

Figure 1.1. Organ shortage in transplantation [3] ...................................................................................... 1
Figure 1.2. Dissertation structure ........................................................................................................... 5
Figure 2.1. An example of pig livers of different conditions .................................................................... 7
Figure 2.2. Liver biopsy data collection .................................................................................................. 12
Figure 2.3. Significant variables selected in models ................................................................................. 17
Figure 3.1. A schematic of the proposed framework ............................................................................... 27
Figure 3.2. Single liver logistic regression model predictor selection ..................................................... 32
Figure 3.3. Four livers overall MR-MTL logistic regression model goodness of fit ................................. 33
Figure 3.4. Four livers overall MR-MTL logistic regression model predictor selection ....................... 34
Figure 3.5. Leave-one-liver-out MR-MTL logistic regression model goodness of fit evaluation 34
Figure 3.6. Leave-one-liver-out MR-MTL logistic regression model predictor selection .................. 36
Figure 3.7. Liver #1 PC loading graphs .................................................................................................... 38
Figure 4.1. Liver preservation .................................................................................................................. 47
Figure 4.2. Image data ............................................................................................................................ 48
Figure 4.3. Image pooling process and result of Liver #1 ....................................................................... 49
Figure 4.4. Prediction performance comparison ...................................................................................... 54
Figure 4.5. Liver #1 variable selection results .......................................................................................... 55
Figure 4.6. Variable selection results for STSVS at $T = 2$ .................................................................... 57
Figure 4.7. Variable selection results for (a) Liver #2, (b) Liver #3 $T = 2$ (c) Liver #4 at $T = 3$ 58
Figure 5.1. Kidney viability experiment .................................................................................................. 69
Figure 5.2. Spatial-temporal biopsy score response surfaces ................................................................. 71
Figure 5.3. Needle insertion experiment .................................................................................................. 72
Figure 5.4. Forecasting accuracy (RMSE) of kidney viability case .......................................................... 75
Figure 5.5. “Cold start” predicted biopsy score and actual biopsy score comparison ............................. 76
Figure 5.6. Predicted biopsy score and actual biopsy score comparison at the last timestamp ............... 76
Figure 5.7. Task similarities by Euclidean distance of model coefficients .............................................. 77
Figure 5.8. Forecasting accuracy (RMSE) of needle biopsy case ............................................................ 79
Figure 6.1. The proposed decision-making flow for UNOS ................................................................... 84
Figure A1. Pearson residual analysis of each liver in logistic regression models .................................... 93
Figure A2. Residual Analysis .................................................................................................................. 94
Figure A3. Leave-one-liver-out MR-MTL logistic regression model residual analysis .......... 95
Figure A4. Leave-one-liver-out MR-MTL logistic regression model goodness of fit evaluation 96
Figure B1. Pearson residual analysis of each liver in logistic regression models ................. 99
Figure B2. Four livers overall MR-MTL logistic regression model residual analysis .......... 100
Figure B3. Leave-one-liver-out MR-MTL logistic regression model residual analysis .......... 101
Figure C1. Liver #2 PC loading graphs ........................................................................ 102
Figure C2. Liver #3 PC loading graphs ........................................................................ 103
Figure C3. Liver #4 PC loading graphs ........................................................................ 104
Figure D1. Image pooling results for (a) liver # 2 (b) liver #3 (c) liver #4 ...................... 105
Figure F1. Liver #1 variable selection results for STSVS at (a) $T=2$, (b) $T=3$, (c) $T=4$, and (d)
$T=5$ .................................................................................................................................. 107
Figure G1. Pairwise distance between model coefficients in Case Study 1 ...................... 108
Figure G2. L1O pairwise distance between model coefficients in Case Study 1 .............. 109
Figure G3. L2O pairwise distance between model coefficients in Case Study 1 .............. 110
Figure G4. L3O pairwise distance between model coefficients in Case Study 1 .............. 110
Figure G5. L4O pairwise distance between model coefficients in Case Study 1 .............. 111
Figure G6. L5O pairwise distance between model coefficients in Case Study 1 .............. 111
Figure G7. L6O pairwise distance between model coefficients in Case Study 1 .............. 112
Figure G8. Pairwise distance between model coefficients in Case Study 2 .................... 113
Figure G9. Pairwise distance between model coefficients in Case Study 2 (16 markers) .. 114
Figure H9. Future research discussions ........................................................................... 115
List of Tables

Table 2.1. Predictor variables and quality responses ................................................................. 12
Table 2.2. Prediction results ........................................................................................................ 16
Table 3.1. Four livers overall MR-MTL logistic regression model classification error .......... 33
Table 3.2. Leave-one-liver-out MR-MTL logistic regression model classification error .......... 35
Table 5.1. Kidney locations and biopsy scores summary ............................................................. 70
Table A1. Single liver logistic regression model classification error ........................................... 92
Table A2. Four livers overall MR-MTL logistic regression model classification error ........... 93
Table A3. Leave-one-liver-out MR-MTL logistic regression model classification error .......... 94
Table A4. Classification errors comparison between the full data set and the truncated data set 97
Table E1. Prediction performance of Liver #1, Liver #2, Liver #3, and Liver #4 ......................... 106
Chapter 1. Introduction

1.1. Motivation

Organ transplantation involves removing organs from a donor and transplanting them into the recipients, and it is an effective therapy for patients with end-stage organ failure. In practice, successful organ transplantations save many patient lives and bring them return to normal daily activities [1]. However, the organ shortage is a critical issue since the demand far exceed supply [2]. Figure 1.1 shows that 113,759 people in the United States were on the waiting list by the end of 2018 compared to only 17,554 US organ donors in the same year. As a result, only 36,529 organs transplantations were performed in 2018 [3]. Moreover, about 20% of the transplantable organs are discarded every year, which is a huge waste in the context of organ scarcity [4]. This is primarily due to the preservation inefficiency and lack of methods to assess the viability of organs without damaging tissues.

Figure 1.1. Organ shortage in transplantation [3]
Visual inspection applied by surgeons and biopsy performed by pathologists are most frequently used assessment methods. However, visual inspection is subjective based on surgeons’ personal experience and training [5]. Biopsy evaluation is examining extracted sample cells or tissues under microscope to determine the presence of a disease or viability of an organ. Compared with visual inspection, biopsy evaluation is more objective and can provide more evidential facts to support the evaluation results [6]. On the other hand, the invasive nature of biopsy practice will cause permanent damages on organ tissues [7], which will potentially reduce the utilization of donor organs. In addition, the lack of viability forecasting methods limits the organ matching arguably find the most suitable recipient [8]. Therefore, quantitative evaluations in real-time based on non-invasive methods are in great needs to reduce the organ discard rate in order to save more end-stage patients. However, online organ viability assessment is very challenging due to four reasons: 1) the assessment results varies among different surgeons due to inconsistent guidelines and procedures to assess organ viability [7]; 2) there is the lack of evidence in literature to use non-invasive biological measurements to correlate with organ viability; 3) the organs are heterogeneous (e.g. size, donor gender, donor age, et al.) so that is difficult to model the viability due to individual-specific degradation path [8]; 4) most assessment methods can only applied at present time, so how to forecast the viability of similar-but-non-identical organs at a future time is still unknown. These challenges provide me the motivation to create non-invasive data-driven methods to predict and forecast the organ viability in real-time. Therefore, a comprehensive literature review is conducted to solve these challenges.

1.2. State-of-the-Art

With the increasingly perfecting of transplantation related hardware and medical skills, organ transplantation develops rapidly since mid-twentieth century and becomes a popular research area
Some assessment methods have been studied in literature, such as pre-transplant screening, which requires both donor and recipient to receive the infectious disease screening [10]. However, it is not always realistic because some of the donor organs are harvested immediately after accidental deaths but it takes 2 to 3 days to generate a screening report. In such a situation, those tests are impossible to conduct due to limited time. Thus, there is a pressing need to find effective and informative methods with a scientific platform that can be correlated to organ viability in real-time. A machine perfusion system (MPS) has been developed to keep organs alive outside the body while preserving functions, and it significantly reduce the risk of delayed organ functions [11, 12]. MPS is a reliable platform that can generate and record rich in situ data, such as pressure, PH value, etc., because it has a closed-loop system to mimic blood circulation with real-time data collection sensors [12]. However, MPS does not have the function to assess the organ viability. The infrared (IR) camera is an ideal solution to assess organ viability due to its non-invasive nature and temperature sensitivity to bodily dysfunction [13]. Although infrared equipment has been used for medical diagnosis in recent years, it has not been related to organ viability diagnosis. Based on the development of hardware, we are able to collect data to investigate new biological measurements, which potentially have strong correlations with organ viability.

Due to its importance, organ transplantation does not only attract attentions from medical doctors but also from other communities, such as engineering and computer science. Many efforts have been made to provide solutions and interpretation to solve transplantation issues from other perspectives. Computer-aided diagnosis (CAD) is developed to assist medical image interpretation due to its high speed, accuracy, and reliability to capture image features [14]. In CAD, machine learning is a mainstream direction in medical diagnosis and disease detection [15]. However, it does not provide medical interpretation because its algorithm is widely known as ‘black box’. In
addition, it also requires a large amount of data to keep training and modifying the model [16]. In organ transplantation, the medical data, especially in large volume, may not be available due to its privacy and confidentiality [17]. As a result, statistical modeling, as an appropriate solution, is performed to evaluate medical test for classification and prediction purposes to handle small data set and provide medical interpretation [18]. The lack of effective data-driven modeling to predict and forecast organ viability in literature ignites the inspiration of this dissertation.

1.3. Research Objectives

The overall objective of this dissertation is to advance the knowledge on modeling organ viability by developing online quantitative and non-invasive methods to predict and forecast the viability of heterogeneous organs in transplantation. It is extremely important because such methods will effectively reduce the organ discard rate. As a result, more end-stage patients will be saved due to significant unrealized transplantation potential existing among discarded organs. The proposed data-driven organ assessment methods can quantitatively provide real-time prediction and forecasting of the viability during the preservation stage without damaging the organs. To investigate the overall objective, Figure 1.2 demonstrates an illustration of the structure of this dissertation. Four specific objectives are listed:

1) to create a method that can objectively evaluate biopsy images to validate the biopsy results by jointly modeling the quantitative and qualitative responses with extracted biopsy imaging features.

2) to correlate the high-dimensional infrared data and the liver viability by extracting features from infrared images to predict liver viability with a multitask learning logistic regression model.

3) to increase the accuracy of predicting the liver viability based on spatial-temporal infrared data by spatial-temporal smooth variable selection without feature engineering.
4) to forecast the kidney viability based on historical data by multitask general path model to support kidney matching decision. To validate the generality of the proposed method, a needle biopsy case study is used to improve needle insertion accuracy.

By investigating these four objectives, these non-invasive and quantitative viability assessment methods are completed to evaluate the organ viability. First, Chapter 2 proposes QQ models that can validate biopsy evaluation results by quantifying the correlation between extracted biopsy image features and liver viability. Second, Chapter 3 introduces a non-invasive and online solution to assess the liver viability without damaging the liver tissue by biopsy. Third, Chapter 4 further investigates the infrared imaging data by considering not only spatial but also temporal correlations to improve the liver viability predicting accuracy without feature engineering. In addition, medical interpretation is provided by highlighting the significant regions for predicting liver viability. Fourth, knowing the historical data of viability loss path of each kidney, the kidney viability can be forecasted by a multitask general path model.

Figure 1.2. Dissertation structure
1.4. Organization of the Dissertation

The rest parts of this dissertation are organized as follows. In Chapter 2, a preliminary study of quantitative and qualitative models to jointly predict porcine liver viability is proposed based on features from biopsy images to validate the biopsy results. This chapter is based on an earlier published paper in IIE Annual Conference 2015 [7]. In Chapter 3, a multi-task learning logistic regression model is proposed to assess the cross-liver viability by correlating liver viability with spatial infrared data by applying principle component analysis to extract infrared imaging features validated by porcine livers during preservation. This chapter is based on an earlier published paper in Computer Methods and Programs in Biomedicine [19]. In Chapter 4, a patio-temporal smooth variable selection method is proposed to predict liver viability by considering both spatial and temporal correlation in modeling without feature engineering, which is also validated by infrared data of porcine livers. In Chapter 5, a multitask general path model is proposed to forecast kidney viability during preservation by transferring knowledge from learning the commonality of all kidneys and the heterogeneity of each validated by porcine kidney. This forecasting method is also generalized to apply to needle biopsy tissue deformation case study with the objective to improve the needle insertion accuracy. In Chapter 6, the conclusions of this dissertation are summarized. Moreover, the future research directions are discussed in the end.
Chapter 2. Quantitative and Qualitative Evaluation for Organ Preservation in Transplant

2.1. Introduction

Organ transplants help many patients to prolong their lifetime. For example, there are approximately 4500 liver transplants performed every year. In addition, the number continues to grow, with a 10-year survival rates of more than 53% [20]. Although the demand for organs is high, fewer than 20% of patients who need a transplant actually receive one [21]. This is primarily due to the shortage of available and transplantable organs. Many potential transplants are simply thrown away before use because organs are not effectively preserved or quality is subjectively considered insufficient for use.

Currently, organ evaluations are mainly performed based on noninvasive visual inspection by surgeons and procurement personnel, or invasive biopsy image evaluations by pathologists. Both types are subjective judgments. Thus, these evaluations are likely to be inconsistent. For example, Figure 2.1 shows pictures of three livers (pig livers) with different conditions preserved in our lab. Figure 2.1 (a) shows a viable liver right after the harvest. Figure 2.1 (b) shows a failed liver, but
it has similar appearance as a viable liver. Figure 2.1 (c) shows a failed liver with obvious deterioration. It is clear that by only comparing their appearance, it is hard to distinguish the viable livers from the failed ones by visual inspection, even if the evaluators are experienced personnel. So far, the most accurate way to evaluate organs viability for transplant is the invasive biopsy, though too many biopsies may damage the organ tissue and affect its functionality. Biopsy is a widely used medical test consisting of sampling cells or tissues for microscopic examination [22].

In liver transplant, the biopsy removes the tissue, usually from the tip of a lobe. The removed tissue is preserved, sectioned, and stained with dyes. The biopsy sections are then evaluated microscopically by experienced pathologists to determine if the liver is viable for the transplant surgery. A binary (qualitative) evaluation helps the surgeon to determine if the transplant surgery should proceed. At the same time, the number of dead cells as a quantitative evaluation in the biopsy images is also analyzed and counted, which quantifies the degree of damage in livers in the preservation and transportation stage. This is especially important for those livers considered as nonviable to assist determining root cause diagnosis of dead cells. Although both the binary evaluation and the qualitative response are used to evaluate livers, they are mainly performed by pathologists in a stressful environment, potentially leading to inaccurate results. For example, to prevent livers from extensive ischemic damage, the transplant surgery needs to start within a roughly eight-hour time interval after the procurement. Thus, pathologists need to evaluate the organ at any time (2:00 am or anytime during the night) and provide a “go/no go” decision, to meet that challenging scheduling in many transplant cases.

Therefore, this paper focuses on developing an objective measure of the organ quality to assist the pathologist in a less stressful environment to determine organ transplant quality. A statistical model, called “QQ models”, is used to simultaneously predict the binary evaluation whether the
organ is viable, and the continuous response, the number of dead cells for further diagnosis. A QQ model is a group of regression models for both Quantitative and Qualitative response variables [23]. In this organ evaluation problem, the predictors will be the biopsy image features to quantify the cell morphology. A detailed summary of these predictors will be discussed in Section 3. With the QQ models, the evaluation whether the organ is viable, and the number of dead cells can be predicted by using objective biopsy image features. In this way, the prediction model provides an objective evaluation to assist pathologists’ subjective evaluation. The QQ models also investigate the correlation of the significant biopsy image features and pathologists’ evaluation, thus improve the understanding of the significant factors for decisions made by pathologists. The model has potential to enhance pathologist training and performance monitoring for organ viability evaluation.

The paper is organized as follows. In Section 2.2, we review the literature on the advancement of biopsy techniques in medical applications, and corresponding statistical analysis based on biopsy image data. In Section 2.3, we discuss the proposed methodology, including image feature extraction, QQ models, and model selection. The proposed method is used in a real case study based on pig liver biopsy images in Section 2.4. Both the prediction performance and the significant predictors of the QQ models are evaluated. Finally, the conclusion is drawn and the future work is discussed in Section 2.5.

2.2. Research Background

The biopsy technique can be traced back to the 11th century, when Ab al-Kasim reports therapeutic puncture of thyroid [6]. Diamantis et al. (2009) provided a review of history for fine-needle aspiration biopsy [6]. However, the biopsy becomes more significant in the 20th century with the use of microscope. Dudgeon and Patrick (1927) used the biopsy of tumors for microscopic
diagnosis [24]. However, the procedure of the biopsy can only select a small number of tissues, which may not reflect the condition of large organs and their functionality. In addition, the procedure of biopsy is not standardized. Depending on the technique of the practitioners, large Type I and Type II errors may be encountered. To overcome this, Zajicek et al. (1974) proposed biopsy and diagnostic criteria in different conditions [25]. The use of biopsy technique was introduced to organ transplant in 1980s. A research group lead by D’Alessandro proposed conducted biopsies and predicted liver viability in transplant [26]. Data from 1987 to 1990 had been collected to validate this proposal [26]. Based on the data, the practitioners can accurately evaluate livers from donors. The outcome of biopsy evaluation was shown to significantly correlate with the patients’ survival rate after liver transplant. In 2000, a new three-dimensional ultrasound guided biopsy was being introduced for breast cancer diagnosis [27]. The researchers found the method a reliable tool for fine needle biopsy to reach desired target region. A scoring system for muscle biopsy evaluation was also used for juvenile dermatomyositis, which assisted the practitioners to improve the accuracy of their subjective evaluation, based on statistical analysis of biopsy images [28].

The investigation of biopsy images for medical applications by using quantitative methods has received many attentions in recent years. Chevallie et al. (1994) proposed a semi-quantitative scoring system for hepatic fibrosis in needle liver biopsy specimens [29]. Kim et al. (1997) used testis biopsy images to quantify spermatogenic cell type [30]. O’Brien et al. (2000) used fibrosis ratio as a feature variable to evaluate liver biopsy specimens with Chronic Hepatitis C [31]. Masseroli et al. (2000) used image analysis and studied the correlation of an image feature based score and the semi-quantitative indexes of fibrosis [32]. Campos et al. (2014) proposed image processing methods to construct fibrosis indexes and compared them with benchmark methods for
fibrosis quantification in Hepatitis C [33]. In summary, all these researchers focused on developing a quantitative index based on biopsy image features to assist the evaluation and diagnosis in their medical applications. However, in organ transplant, both quantitative and qualitative evaluations are vital. The qualitative evaluation to determine the organ’s viability has captured the pathologists’ knowledge for judgment, while the quantitative evaluation on number of dead cells is important to identify root causes for organ failure during preservation. In the literature, modeling both quantitative and qualitative evaluations has not been reported so far. The lack of statistical models for both types of responses motivated us to investigate this research problem.

2.3. Proposed Methodology

2.3.1. Biopsy Procedure and Data Processing

To consistently evaluate organs at different conditions, all the biopsies are performed at the same organ to avoid the heterogeneity over different organs. The organs are preserved in the same environment (e.g., temperature), and the biopsies are performed at different locations and different time. Figure 2.2 (a) shows the location and time for the biopsies of a liver. The locations of these five biopsies are marked as cylinders. These biopsies are taken at 0, 4, 8, 12, 24 hours after harvest. At each time point, a group of five biopsies for the same lobe are taken, thus it avoids the damages to other lobes. In this way, 25 biopsy specimens have been conducted, with tissues from a natural degradation process of the liver driven by time in a controlled lab environment. Each biopsy specimen was magnified 100 times, and four images from the same biopsy specimen have been taken to increase the sample size for this analysis. To ensure the consistence of different image samples, the images have the same resolution. In total, 100 biopsy images were collected. Figure 2.2 (b) shows an example of the biopsy images.
Figure 2.2. Liver biopsy data collection
(a) biopsy locations and time for the same liver (b) an example of biopsy image from the liver in (a)

Table 2.1. Predictor variables and quality responses

<table>
<thead>
<tr>
<th>Type</th>
<th>Notation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image feature variables</td>
<td>$x_1$: continuous</td>
<td>Number of nuclei</td>
</tr>
<tr>
<td></td>
<td>$x_2$: continuous</td>
<td>Number of round nuclei</td>
</tr>
<tr>
<td></td>
<td>$x_3$: continuous</td>
<td>Number of large nuclei</td>
</tr>
<tr>
<td></td>
<td>$x_4$: continuous</td>
<td>Number of nuclei without hole</td>
</tr>
<tr>
<td></td>
<td>$x_5$: percentage</td>
<td>Non-white to whole area ratio</td>
</tr>
<tr>
<td></td>
<td>$x_6$: continuous</td>
<td>Number of cells with normal shape</td>
</tr>
<tr>
<td></td>
<td>$x_7$: continuous</td>
<td>Number of deteriorating cells</td>
</tr>
<tr>
<td></td>
<td>$x_8$: continuous</td>
<td>Number of deformed cells</td>
</tr>
<tr>
<td></td>
<td>$x_9$: percentage</td>
<td>White area percentage in the photo</td>
</tr>
<tr>
<td></td>
<td>$x_{10}$: continuous</td>
<td>Number of cell packing (Eccentricity)</td>
</tr>
<tr>
<td></td>
<td>$x_{11}$: percentage</td>
<td>Cell vacuolation (holes)</td>
</tr>
<tr>
<td>QQ responses</td>
<td>$y$: continuous</td>
<td>Number of dead cells</td>
</tr>
<tr>
<td></td>
<td>$z$: binary</td>
<td>Binary indicator for organ viability</td>
</tr>
</tbody>
</table>

Based on the biopsy image, features were extracted through Matlab® image processing toolbox. At the same time, a pathologist received all biopsy images. Following the standard procedure to subjectively evaluate the images, the pathologist reviewed the images without knowing the location and the time of the image collection. He determined whether the biopsy images come from a viable tissue sample, or a failed one, which was then used as a qualitative response. He also provided an estimate of dead cells in these images, which will be used as a quantitative response in the later analysis. Table 1 summarized the notation and meaning of the collected image feature variables and corresponding response variables. Because all the image feature variables are
objective measures, which can be easily obtained from the image processing methods, the objective of the statistical analysis is to use the image feature variables to predict the subjective quantitative and qualitative (QQ) responses.

2.3.2. Quantitative and Qualitative Models

In the literature, there is limited work for jointly modeling the QQ responses. Fitzmaurice and Laird (1997) investigated the relationship for QQ responses. However, the binary response is predicted in a marginal logistic regression, and the quantitative response is predicted in a marginal linear regression model, which may not have accurate predictions [23, 34]. In this paper, we adopted the “QQ models” to consider both the quantitative and qualitative responses in the quality control process [23]. This joint modeling keeps the integrity of the two types of responses within one framework. The advantage of the QQ models is that they consider the association between the two types of responses. The difference of this paper and our earlier work in QQ models [23] lies in that we first use a logistic regression model to predict the binary response, and then use a linear regression model to predict the quantitative response. Thus, the models are estimated sequentially in this paper, rather than estimated simultaneously as shown in [23]. This step guarantees that the modeling performance for the binary response is not compromised, because the likelihood function is maximized in parameter estimation in separate modeling, while the simultaneous modeling may not yield the maximal likelihood estimates of model parameters. This is important for the organ evaluation application, because the evaluation for a viable organ or a failed one is the more critical question to answer than the prediction of dead cells. We denote the qualitative response (a binary variable) as \( z \), the quantitative response (a continuous variable) as \( y \), and the predictor vector as \( x = (x_1, \ldots, x_p)' \) with \( p \) predictors. The QQ models are estimated as follows.

First, a logistic regression model is used for modeling the binary response,
\[ z_i = \begin{cases} 1, & \text{w.p. } P(x_i) \\ 0, & \text{w.p. } 1 - P(x_i) \end{cases} \quad \text{with } (x_i) = \frac{\exp(x_i'\eta)}{1+\exp(x_i'\eta)}, \quad (2.1) \]

where \( z_i \) is the binary response for the \( i \)-th sample; \( x_i \) is the predictor vector of the \( i \)-th sample; and \( \eta = (\eta_1, \ldots, \eta_p)' \) is a vector of logistic regression parameters. \text{w.p.} stands for “with probability”.

Second, two conditional linear regression models based on the binary response are used for modeling the quantitative response,

\[ y_i | z_i = 1 \sim N(x_i'\beta^{(1)}, \sigma_1^2), \quad (2.2) \]
\[ y_i | z_i = 0 \sim N(x_i'\beta^{(2)}, \sigma_2^2), \quad (2.3) \]

where \( y_i \) is the quantitative response for the \( i \)-th sample; \( \beta^{(m)} = (\beta_1^{(m)}, \ldots, \beta_p^{(m)})', \) \( m=1, 2 \) are the corresponding regression parameters; \( \sigma_1^2, \sigma_2^2 \) are the residual variance for the two linear regression models. \( y_i | z_i = 1 \) and \( y_i | z_i = 0 \) follow two normal distributions with different regression means and residual variances.

To balance the trade-off between model complexity and prediction performance, the Bayesian Information Criterion (BIC) was adopted for model selection [35], first for the logistic regression model, then the linear regression models. The BIC score can be calculated as

\[ BIC = -2L + q \log(n), \quad (2.4) \]

where \( L \) is the log-likelihood function for the logistic regression model or the linear regression models, \( q \) is the number of significant predictors, and \( n \) is sample size of the training data.

To predict the QQ responses, the binary response \( z_i \) is predicted first as

\[ \hat{z}_i = \begin{cases} 1, & \text{w.p. } \hat{P}(x_i) \\ 0, & \text{w.p. } 1 - \hat{P}(x_i) \end{cases} \quad \text{with } \hat{P}(x_i) = \frac{\exp(x_i'\hat{\eta})}{1+\exp(x_i'\hat{\eta})}, \quad (2.5) \]

where \( \hat{z}_i, \hat{P}(x_i) \) are the corresponding predicted values for the \( i \)-th new observation; and \( \hat{\eta} \) is the estimated parameters in the logistic regression model with the BIC for variable selection.

Based on the values of \( \hat{z}_i \), the quantitative response \( y_i \) will be predicted as
\[ \hat{y}_i = \begin{cases} x'_i \hat{\beta}^{(1)}, & \text{if } \hat{z}_i = 1 \\ x'_i \hat{\beta}^{(2)}, & \text{if } \hat{z}_i = 0 \end{cases} \]  

(2.6)

where \( \hat{y}_i \) is the predicted value for the \( i \)-th new observation; and \( \hat{\beta}^{(1)} \) and \( \hat{\beta}^{(2)} \) are the estimated parameters in the linear regression models with the BIC for variable selection.

### 2.4. Case Study

Recalling the data collection procedure in Section 2.3.1, the original data were preprocessed. For the binary response, we used “1” for viable organs, and “0” for failed organs. We also standardized the predictors and centralized the quantitative response variable by subtracting the mean, because the mean and standard deviation of these variables are very different. We didn't standardize the quantitative response variable because we would like to have the variation estimated in the residuals of the linear regression models. For the QQ models, we used \( x_1, \ldots, x_{11} \) and their two-way interactions as predictors, and these variables are denoted in Table 2.1. In total, there are 66 predictors in the full model. We also performed residual analysis to validate the assumptions, and removed the outliers. After the outlier removal, we used 92 samples for modeling. To evaluate the modeling performance, we investigated significant predictors and compared the prediction performance of the QQ models with traditional regression methods in 10-fold cross validations (CV).

#### 2.4.1. Prediction Performance

Table 2.2 summarizes the prediction results of the average prediction errors of 10-fold CV for training data and testing data. To predict the binary evaluation whether the organ is viable or not, we summarize the Type I error, Type II error and overall misclassification error. For the testing data, the overall misclassification error based on the logistic regression model is 9.7%, and Type I error and Type II error are 11.7% and 8.3%, respectively. By checking with the pathologist, the prediction result is acceptable. After the classification of the specimens, we can divide the
specimens into two groups. Then we use the QQ models to predict the quantitative response. We found that the prediction error of $y|\hat{z} = 0$ is 0.551 for the testing data, meaning the prediction error is 0.551 cells on average, which is much smaller than the prediction error by using the benchmark linear regression models (0.957). But, if we consider the QQ models as a whole to predict $y$, the prediction error is 1.363 for the testing data, which is larger than that of the benchmark model. This is because the prediction for $y|\hat{z} = 1$ is poor with a testing error as 1.544. We further investigate the reason, and found that when the organs are predicted as good ($\hat{z} = 1$), the number of dead cells is very small. Thus, $y|\hat{z} = 1$ has consistently small values of $y$ in the training data, leading to a poor estimation of the model for $y|\hat{z} = 1$. However, the number of dead cells is more important for bad organ assessment, i.e., the prediction of $y|\hat{z} = 0$ is more important. We found that the model is still acceptable since the QQ models can have much better prediction for $y|\hat{z} = 0$ comparing the benchmark linear regression model.

Table 2.2. Prediction results

<table>
<thead>
<tr>
<th>Responses</th>
<th>Prediction</th>
<th>Training Errors</th>
<th>Testing Errors</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{z}$</td>
<td>Overall</td>
<td>1.9%</td>
<td>9.7%</td>
</tr>
<tr>
<td></td>
<td>Type I</td>
<td>4.0%</td>
<td>11.7%</td>
</tr>
<tr>
<td></td>
<td>Type II</td>
<td>0.8%</td>
<td>8.3%</td>
</tr>
<tr>
<td>$y$</td>
<td>QQ $y</td>
<td>\hat{z} = 0$</td>
<td>0.688</td>
</tr>
<tr>
<td></td>
<td>QQ $y</td>
<td>\hat{z} = 1$</td>
<td>0.906</td>
</tr>
<tr>
<td></td>
<td>QQ overall</td>
<td>0.843</td>
<td>1.363</td>
</tr>
<tr>
<td></td>
<td>Linear Regression</td>
<td>1.039</td>
<td>0.957</td>
</tr>
</tbody>
</table>

2.4.2. Variable Selection

To understand the correlation between pathologist’s evaluation and image features and the correlation between the number of dead cells and image features, we also investigated the significant predictors identified in all models. Figure 2.3 summarizes the significance of predictors, where each row and each column represent predictors from $x_1$ to $x_{11}$, defined in Table 2.1. The
diagonal elements represent the main effect of the variables, and the off-diagonal element at the $i$-th row and the $j$-th column represents the two-way interaction $x_ix_j$, $i = 1, ..., 11, j = 1, ..., 11$. If the predictor is significant, then the corresponding element is highlighted in black color.

![Images](image1.png)

**Figure 2.3.** Significant variables selected in models
(a) logistic regression model (b) benchmark linear regression model
(c) $y|\hat{z} = 1$ in QQ models (d) $y|\hat{z} = 0$ in QQ models

Based on Figure 2.3 (a), we conclude that the main effects of the number of nuclei without hole ($x_4$), the non-white to whole area ratio ($x_5$) and the number of cells with normal shape ($x_6$), and some other interactions are significant predictors to determine if the specimen comes from a viable organ, or a failed one. By comparing Figure 2.3 (b), (c) and (d), we conclude that the significant predictors for $y|\hat{z} = 1$ and $y|\hat{z} = 0$ are very different, which implies that the significant predictors to predict the number of dead cells are changing given that the organ is viable or failed. Therefore, it is necessary to model the quantitative response given different values of the qualitative evaluation. Note that the significant predictors for $y|\hat{z} = 0$ are the main effects of the
number of cells with normal shape ($x_6$), the number of deteriorating cells ($x_7$) and some other interactions, which is very meaningful to interpret.

2.5. Conclusion and Future Work

Organ transplantation is an important way to prolong the lifetime of patients with organ failure. However, the current organ evaluation methods are based on either subjective visual inspection or biopsy image evaluation. Neither of methods are effective. In this paper, we propose to use a QQ model to jointly predict the binary response (whether the organ is viable), and the quantitative response (the number of dead cells). In general, the prediction performance for the binary response is promising, while the prediction for the number of dead cells of failed organs is better than the benchmark regression models. Based on the QQ models, significant predictors are identified and the significant predictors for $y|\hat{z} = 1$ and $y|\hat{z} = 0$ are different, indicating different correlation structures of the variables for viable organs and failed organs. The QQ models provide an objective evaluation based on biopsy image features, which have its potential applications for pathologist training and performance monitoring.

This paper is a pioneering work for modeling both quantitative and qualitative evaluations in organ transplants. Therefore, there are still a lot of future work that needs more investigation, mainly in two directions. First, the improvement on statistical analysis should be made. In this paper, the QQ models are used for a single qualitative response, and a single quantitative response. How to extend this work to multiple responses calls for more efforts. The model assumptions should also be generalized for multi-level categorical variables, count data, or nonlinear model structures. For the quantitative response, advanced models could also be incorporated such as data mining models, nonlinear models to address the nonlinear effects of the predictors to the response. Although the QQ models in this paper predict number of dead cells, which is a count variable, the prediction
performance with the normal residual assumption is reasonably well. Second, the planned experiments should consider additional factors to affect modeling performance, such as having multiple pathologists’ evaluation at multiple times. In this way, we can assess the agreement of the evaluation results for the same pathologist at different times, or among different pathologists. The size of the organ also affects the biopsy results. Therefore, experiments considering spatial effects of the biopsy locations will be studied. This study only focuses on one organ, therefore, multiple organs should be considered to model the heterogeneity of organs from different donors.
Chapter 3. Non-invasive Assessment of Liver Quality in Transplantation based on Thermal Imaging

Analysis

3.1. Introduction

End-stage liver disease claims 25,000 lives in the U.S. every year. For those with end-stage organ failure, liver transplantation is the single most effective way to prolong patients’ life. Liver transplants have an about 85% to 90% one-year survival rate [36]. However, it is usually hard to find a viable liver to transplant, since demand far exceeds supply. In August, 2016, there were 14,633 candidates in the national waiting list for liver transplantation, but fewer than 20% of patients will actually receive one [37], based on past experiences. Moreover, the survival rate of liver transplants highly depends on the rapid liver preservation and liver quality evaluation, which remain challenging issues. Many potentially “transplantable” livers are discarded before implantation due to the preservation “hardware” inefficiencies (e.g., current preservation methods only allow preservation for <12 hours) and the lack of accurate liver quality prediction models (e.g., visual inspection, currently used, is highly subjective) [38]. The most commonly used preservation method is to place explanted livers in ice-filled Styrofoam containers [39]. Such an approach can reduce the liver metabolism rate, but cannot minimize the ischemic cellular damage initiated during organ procurement [40].

The accurate evaluation of liver quality is a very challenging issue. A first step in the current evaluation process is subjective visual inspection of the liver surface, which heavily depends on the evaluator’s experience and expertise, and thus suffers from potential evaluation bias. A second
step is procurement and histologic evaluation of a small (<5 mm diameter) biopsy. Biopsies can produce useful information (depending on the skill in procurement and interpretation of the morphologic features). However, taking biopsies can cause focal damage to the liver, and, depending on the site sampled, may not represent the histomorphology of the entire liver [41]. Due to the limitations of current preservation and evaluation methods, many viable livers (about 50% of potential donations) are not used for liver transplantation [4].

In recent years, machine organ preservation systems, that emulate physiologic circulation, have been used to perfuse explanted livers and to conduct engineering studies [42]. The advantages of the liver preservation systems include: (1) providing physiologically-relevant, regulated waveforms and fluid circulation for organ preservation; and (2) extending the lifetime of large and complex organ such as livers [43]. However, these systems do not provide real-time, objective metrics for assessment of liver viability. In this paper, we describe a process for liver viability evaluation and for monitoring tissue health over time by using real-time data collection and analysis. A non-invasive real-time monitoring system, with high-resolution infrared (IR) imaging camera and computational thermography were used to capture the liver surface temperature during preservation, in order to perform viability analysis. The use of this non-invasive system prevents potential tissue damage incurred during biopsy procedures and reduces the time needed for collection and evaluation of biopsies on livers.

Using IR images and computational thermography for liver viability evaluation is non-trivial. First, the high dimensional surface temperature data (i.e., the IR images) poses a great challenge for building an efficient and interpretable evaluation model. The value and applicability of useful feature extraction from the high dimensional surface temperature data, as liver quality evaluation modeling predictors, is an open question. In this paper, dimensional reduction methods, such as
Principal Component Analysis (PCA) [44-46] will be applied to IR image data to obtain evaluation predictors. Second, we performed a limited number of liver perfusion experiments as a pilot study to validate our approach. The heterogeneity of individual livers in the limited data set presents a major challenge for liver evaluation. In this paper, we use the multi-task learning (MTL) approach [47] to take advantage of the similarity of livers from different donors. In the case study, it is found that the proposed non-invasive liver assessment approach could establish a cost-effective method with an accurate prediction performance. We use our approach to test the hypothesis “high-resolution infrared imaging and computational thermography can be used to assess explanted liver perfusion and effects of perfusion on tissue viability”.

3.2. Research Background

Organ collection and preservation procedures play an important role in maintaining the viability and functionality of explanted organs destined for transplantation [48]. As early as 1930s, Carell invented a sterile perfusion pump to maintain organs outside of human body, leading the way for organ transplantation procedures that were successfully conducted in practice in 1960s [49]. Carell’s method used low temperature solutions to flood the organ, but maintaining organ sterility was challenging [50]. Persufflation (PSF, gaseous oxygen perfusion) has been widely studied for organ preservation over a century, but a major barrier for its application is post-implantation embolization caused by ex vivo PSF [51]. The lack of reliable machine perfusion equipment and technical problems with insuring fluid circulation made static cold storage on ice the major solution in organ preservation [52].

The current methods to assess explanted organ quality are biopsy and visual inspection. The biopsy can provide an accurate diagnosis results. Martin and Ellis invented the technique of modern needle puncture in early 20th century [6]. Menghini used this technique for liver evaluation [22].
The biopsy evaluation method was first applied in organ transplantation in the 1980s, and it remains the most reliable and accurate evaluation method to verify the quality of the organ in the pre-implantation stage [53]. However, the biopsy evaluation method suffers from several drawbacks: (1) the evaluation of biopsy results depends on the pathologists’ experience and expertise; and (2) taking biopsies causes minimal but irreversible damage to the organ [5, 54].

Gross visual inspection is widely used to initially assess liver quality (discoloration due to the presence of intrahepatic lipid and cirrhotic scarring are two key discriminate features). Gross inspection is non-invasive, which prevents the liver from being damaged. However, the visual evaluation method depends on the evaluator’s experience and expertise, and suffers from lack of consistent standards [55].

On the other hand, various sensor data, such as blood pressure [56] and cell morphology image [5], have been used to evaluate organ quality. Recently, IR imaging started to receive attentions for organ quality evaluation and in disease diagnosis and treatment because it is non-invasive and sensitive to bodily dysfunction as a skin-temperature indicator [57-59]. This is primarily because the natural emission temperature differentials between adjacent organ structures detected by IR imaging is relevant to organ preservation and for health condition evaluation. Gorbach et al. applied IR imaging to correlate renal surface temperature with underlying renal ischemia, found that the blood flow oscillations can be indirectly measured by IR imaging, and indicate the early critical renal ischemia [60]. Gorbach, et al. also demonstrated the feasibility of assessing porcine kidney quality by IR imaging under pulsatile perfusion conditions [61].

Based on the measurement data, various modeling approaches can be used for organ quality evaluation. For instance, a multivariate logistic regression analysis was used to model the difference between graft liver transplantation and the whole organ transplantation [62]. Cox
regression models were used to identify the key factors that would cause liver failure in the post-transplantation interval [63]. A hierarchical linear regression approach was used to probe the variation in success/failure rates among liver transplantation centers based on the quality of donated organs [64]. The Quantitative and Qualitative (QQ) models were used to jointly predict the heterogeneous quality conditions in organ transplantation [23]. In this paper, one continuous quantitative response representing the number of dead cells and one binary qualitative response representing the organ viability were jointly evaluated by the features extracted from biopsy images. However, the QQ models were only applied to the single liver evaluation, but not the cross-liver evaluation, in which testing liver quality is predicted according to available liver data. To sum up, the previous methods were either lack of effective approaches to evaluate the organ quality in the pre-surgery stage or could cause permanent damage to the organ.

3.3. Materials and Methods

In order to address the aforementioned limitations, this paper aims to test the hypothesis that “high-resolution infrared imaging and computational thermography can be used to assess explanted liver perfusion and effects of perfusion on tissue viability.” Essentially, this is to test if there are strong correlations between infrared image features and tissue viability.

3.3.1. Experimental Planning

Explanted porcine livers were collected to evaluate the proposed imaging and computational methods. Porcine liver was used due to the similarities that it shares with human liver structurally and genetically [65], and its comparable size to human liver. Traditional liver transplantation modeling studies usually used rat liver because of its relatively small size for easier handling in perfusion and preservation procedures [66]. However, the effects of size and vascular complexity are significant and the study of an appropriately sized model can be essential in studying liver
transplantation metrics [67]. For study, livers were collected from pigs sacrificed at an abattoir. Livers were collected on ice within five minutes of death and flushed with cold (4°C) phosphate-buffered physiologic saline within 15 minutes of death, using one-meter gravity infusion into the portal vein and hepatic artery. After flushing, livers were placed on ice and returned to the laboratory for imaging and measurements. Once in the lab, livers were connected via hepatic artery and portal vein to a machine perfusion system. Once connected to the machine perfusion system, cold (4°C) (two livers) or room temperature (22°C) (two livers) phosphate-buffered physiologic saline was infused and a closed loop reperfusion system was established. Physiologically relevant systolic and diastolic pulse pressures, flows, and pulse rate were individually adjusted for the portal vein and hepatic artery. Livers were perfused for 24 hours and IR images were collected for analysis.

To collect IR images, a FLIR A655sc thermal camera (FLIR Systems Inc., Wilsonville, OR) was mounted at a distance of one meter above the superior surface of the perfused liver. The entire surface of the liver was contained in the imaging field and a frame capture rate of one frame/min was actuated. Experimental data from four porcine livers (ID #1, #2, #3, and #4) were collected for this study. For each liver, surface temperature was captured as a (640 × 480) pixel image by IR camera every 1 minute during a 24-hour time period.

3.3.2. Data Cleaning

Due to the high cost of procuring and interpreting biopsies, biopsy collection was limited and the binary response of each observation was generated under the assumption that liver goes unviable after the 8-th hour. This was concluded from our previous study in biopsy-based liver evaluation modeling [5], and was proved in the research that the machine perfusion preservation can keep a porcine liver viable for 8 hours [68]. In [5], biopsies were taken at 0, 4, 8, 12, and 24 hours from
five locations of the same lobe each time, and it was determined by the biopsy images that liver became unviable after 8-th hour. As a continuous research work following [5], it is reasonable to draw the “8-th hour threshold” assumption as the ground truth for this work that the same type of liver (porcine) undergoing similar harvest and perfusion procedures would go unviable around similar time period. To ensure the data accuracy and to eliminate liver-to-liver difference in terms of failure time, we excluded the data 3 hours before and after the 8-th hour which were 6-th through 11-th hour. 9 outliers of the average image temperature due to abrupt temperature change caused by environmental noise were removed from the 1080 observations (1 image/min × 18 hours) for each liver. A total number of 1071 IR images or frames were obtained for each liver. The modeling results based on the full set of data are listed in the Appendix A (1431 observations).

3.3.3. Proposed Method

A schematic of the computational thermography method is illustrated in Figure 3.1. Temperature data, represented by pixels on each IR image, are considered as raw predictors. The IR images were continuously collected online during the preservation process of the organ. PCA [44-46] was then used to extract temperature features (PC scores) from IR images. These PC scores were then used as predictors for modeling the liver quality, in a logistic regression model with Lasso penalty for singe-liver quality evaluation [69]. A penalized multi-task logistic regression model [70] was then used for cross-liver quality evaluation. The objective of this paper is to find the relationship between PC scores and the liver viability. The principles of each analytical model are introduced in the following sub-sections.

Considering the computational efficiency of calculating PC scores, a liver was divided into six regions by its natural partition instead of treating it as an entirety. An additional advantage by doing so is that one can have more predictors to compare and interpret liver-to-liver difference in
the cross-liver evaluation. The six areas are defined as region 1 (left lateral lobe), region 2 (left medial lobe), region 3 (right medial lobe), region 4 (right lateral lobe), region 5 (caudate lobe), and region 6 (bile) as illustrated in the first block (porcine liver) of Figure 3.1 [71]. Denote \( \xi_i = \left( \xi_i^{(1)^T}, ..., \xi_i^{(6)^T} \right)^T \), \( i = 1, ..., n \), as the temperature vector including the temperature information at each pixel (pixel intensities) of the \( i \)-th liver IR image, where \( \xi_i^{(k)} = (\xi_{i,1}^{(k)}, ..., \xi_{i,J_k}^{(k)})^T \) is the temperature vector for area \( k, k=1, ..., 6 \). \( J_k \) is the number of pixels in area \( k \), \( n \) is the number of images (observations) obtained following the temporal order. \( z_i \in \{0, 1\} \) is the binary response for the \( i \)-th observation, where “1” represents a viable liver observation and “0” represents an unviable liver observation. The set of image data with corresponding binary response is denoted as \( (\xi_i, z_i), i = 1, ..., n \).

![Diagram](image)

Figure 3.1. A schematic of the proposed framework
3.3.3.1. Feature Extraction by PCA

The basic idea of PCA is to apply a vector \( \mathbf{a}_j^{(k)} \) (i.e., PC loading) to transform \( \mathbf{x}_i^{(k)} \) to \( \mathbf{x}_{i,j}^{(k)} = \mathbf{a}_j^{(k)T} \mathbf{x}_i^{(k)} \), \( i = 1, \ldots, n, j = 1, \ldots, J_k, k = 1, \ldots, 6 \). \( \mathbf{x}_{i,j}^{(k)} \) is the \( j \)-th PC score in the \( k \)-th region of the \( i \)-th observation. \( \mathbf{x}_i^{(k)} = (x_{i,1}^{(k)}, \ldots, x_{i,J_k}^{(k)})^T \) is a vector of orthogonal PC scores in the \( k \)-th region of the \( i \)-th observation. In general, most of the variance in the data set can be explained by a very small number of PC scores [72]. Let \( p_k \) be the number of PC scores that can explain equal or greater than 95% of the variance in the data set, where \( p_k \leq J_k \). The PC loading \( \mathbf{a}_j^{(k)} \), which is the eigenvectors of the covariance matrix \( \mathbf{C}_i^{(k)} = \mathbf{X}_i^{(k)T} \mathbf{X}_i^{(k)} \), is estimated by solving Eq. (3.1) to extract the maximum variance from the original data matrix. As a result, the first PC score explains the largest variance in the original matrix, the second PC score explains the second largest variance in the original matrix, and so on.

\[
\hat{\mathbf{a}}_1^{(k)}, \ldots, \hat{\mathbf{a}}_{p_k}^{(k)} = \arg\min_{\mathbf{a}_1^{(k)}, \ldots, \mathbf{a}_{p_k}^{(k)}} \left\| \mathbf{x}_i^{(k)} - \sum_{j=1}^{p_k} \mathbf{a}_j^{(k)T} \mathbf{C}_i^{(k)} \right\|^2
\]

Subject to \( \left\| \mathbf{a}_j^{(k)} \right\| = 1, < \mathbf{a}_i^{(k)}, \mathbf{a}_j^{(k)} > = \begin{cases} 0, & i \neq j, \\ 1, & i = j, \\ \end{cases} \), \( i, j \in (1, \ldots, p_k) \),

where \( \hat{\mathbf{a}}_1^{(k)} \) is the coefficient vector for the 1st PC score in region \( k \), \ldots, and \( \hat{\mathbf{a}}_{p_k}^{(k)} \) is the coefficient vector for the \( p_k \)th PC score in region \( k \). \( \| \cdot \|^2 \) denotes the squared norm and \( < \cdot, \cdot > \) denotes the inner product.

In this paper, the first four PCs for each region take up at least 95% variation of the data, so the first four PC scores were selected as predictors in each region. Therefore, a collection of 24 predictors with 1071 observations (time frames) on four livers constitute the data set. The predictors were standardized prior to the analysis. And the PCA was performed based on all data from the four livers.
3.3.3.2. A Logistic Regression Model for Single Liver Evaluation

In this study, we firstly test the hypothesis that the information extracted by PCA from IR images has a strong correlation with the liver quality within singer liver. This is validated by a regression modeling method. As the most popular linear regression model for binary responses, a logistic regression model [73] was chosen for modeling the single liver viability with PC scores \( x_i^{(k)} \) as predictors, shown as follows,

\[
P(z = 1) = \frac{\exp(\eta_0 + \sum_{k=1}^{p} \sum_{j=1}^{p_k} \eta_j^{(k)} x_{i,l,j}^{(k)})}{1 + \exp(\eta_0 + \sum_{k=1}^{p} \sum_{j=1}^{p_k} \eta_j^{(k)} x_{i,l,j}^{(k)})},
\]

(3.2)

where \( \eta_0 \) is the intercept, \( \eta_j^{(k)} \) \( (j = 1, \ldots, p_k) \) is the regression coefficient of \( j \)-th PC score in \( k \)-th region. Due to a fairly large number of predictors and observations, the logistic regression model with Lasso penalty was used to shrink the coefficients of insignificant predictors to zero, and thus provide an interpretable model estimation and a good classification performance [69]. The objective function of the applied model is shown as follows,

\[
\arg\min_{(\eta_0, \eta) \in \mathbb{R}^{p_k+1}} \sum_{i=1}^{n} \log(1 + \exp(-z_i(\eta_0 + x_i^T \eta))) + \lambda \| \eta \|_1,
\]

(3.3)

where \( x_i = \left( x_i^{(1)}^T, \ldots, x_i^{(6)}^T \right)^T \) is model inputs composed of PC scores, \( \eta = \left( \left( \eta_1^{(1)}, \ldots, \eta_{p_1}^{(1)} \right), \ldots, \left( \eta_1^{(6)}, \ldots, \eta_{p_6}^{(6)} \right) \right)^T \) is a vector of model coefficients, and \( \| \cdot \|_1 \) denotes the \( \ell_1 \)-norm. The objective function was solved by adopting cyclical coordinate descent algorithm, which was fulfilled by using the Glmnet package for MATLAB [74]. The tuning parameter \( \lambda \) was chosen by 5-fold Cross Validation (CV) to minimize the CV error [69].

3.3.3.3. A MTL Logistic Regression Model for Cross-Liver Evaluation

To further apply IR images in liver quality evaluation besides assessing the modeling accuracy within one liver, it is important to identify methods to use IR images to predict the viability of a
new liver, based on a historical database of livers preserved. Therefore, the second problem of interest is to test the hypothesis that the IR imaging technology could be utilized in cross-liver evaluation by comparing the testing liver quality with its known responses. To address the cross-liver prediction problem, the MTL approach was adopted [47]. MTL is powerful in dealing with a group of related data sets (tasks) with predictors and responses sharing commonality. In MTL, these responses are learned jointly so that their commonality can be borrowed for improving the model prediction accuracy over separate modeling of each task. Researchers performed MTL in computational biology and biology image processing in order to capture the similarity between different tasks in genome biology study [75]. MTL has also been conducted in studying human response to flu, where the best treatment to each type of flu was identified [76]. Moreover, high dimensional data such as functional magnetic resonance imaging data was analyzed by MTL [77].

In this paper, we adopted the Mean-Regularized MTL (MR-MTL) logistic regression model for fitting the multi-liver data [70].

$$\min_{\mathbf{W}, \mathbf{c}} \sum_{i=1}^{t} \sum_{j=1}^{n_i} \log(1 + \exp(-z_{i,j}(\mathbf{x}_{i,j}^T \mathbf{W}_i + c_i))) +$$

$$+ \rho_1 \sum_{i=1}^{t} \left\| \mathbf{W}_i - \frac{1}{t} \sum_{s=1}^{t} \mathbf{W}_s \right\|_2^2 + \rho_2 \| \mathbf{W} \|_1, \quad (3.4)$$

where $t$ denotes the number of livers (tasks), $z_{i,j}$ is the binary responses of sample $j$ in the $i$-th task, $\mathbf{x}_{i,j} = ((x_{i,j,1}, \ldots, x_{i,j,p_1}), \ldots, (x_{i,j,1}, \ldots, x_{i,j,p_6}))^T$, and $x_{i,j,r}$ indicates the $r$-th PC of sample $j$ in the $i$-th task. $\mathbf{W}$ and $\mathbf{c}$ are the MTL logistic model coefficients, $\mathbf{W} = (\mathbf{W}_1, \ldots, \mathbf{W}_t)$, $\mathbf{W}_i = (W_{i,1}^{(1)}, \ldots, W_{i,p_1}^{(1)}), \ldots, (W_{i,1}^{(6)}, \ldots, W_{i,p_6}^{(6)})^T$, $i = 1, \ldots, t$, $\| \mathbf{W} \|_1 = \max_{1 \leq j \leq t} \sum_{i=1}^{p_1+\ldots+p_6} |W_{i,j}|$, and $\mathbf{c} = (c_1, \ldots, c_t)^T$. Equation (3.4) contains two penalty terms associated with parameters $\rho_1$ and $\rho_2$. The first penalty term regularizes the deviation of each task from the mean. The second penalty term is the $\ell_1$-norm of $\mathbf{W}$. A large value of $\rho_1$ indicates a large similarity shared among tasks and vice
versa; and the function of $\rho_2$ here is same as the turning parameter $\lambda$ in Equation (3.3) to control the overall model complexity. As a special case of MTL, Equation (3.4) is simplified to Equation (3.3) when the task number $t = 1$. Equation (3.4) was solved by using the MALSAR package for MATLAB [78]. $\rho_1$ and $\rho_2$ were chosen by 5-fold CV [69].

3.4. Results of Case Study

3.4.1. Single Liver Evaluation

In the single liver evaluation, 1071 observations of each liver after PC scores extraction were randomly divided into 70% as training data and the rest 30% as testing data. The logistic regression model shows good fitting results on single liver (#1, #2, #3, and #4) evaluation with 0% errors on both training and testing data sets in terms of Type I error, Type II error, and overall error. Type I error here is defined as the percentage of viable liver images being predicted as unviable among all the viable liver observations; Type II error is defined as the percentage of unviable liver images being predicted as viable among all the unviable liver observations; and the overall error is defined as the percentage of viable or unviable liver being predicted as unviable or viable among all the observations. The Pearson residual analysis graphs [79] are shown in the Appendix B. Up to this point, the PC scores extracted from IR images show a strong correlation with the liver viability for single liver quality evaluation.
Figure 3.2. Single liver logistic regression model predictor selection of (a) Liver #1; (b) Liver #2; (c) Liver #3; (d) Liver #4. Each square cell indicates one predictor, following the order of Regions 1 to 6 from left to right, and PC scores 1 to 4 from top to bottom. The black-filled cells indicate the selected significant predictors with non-zero fitted coefficients for evaluating the liver quality, and the white cells are the predictors with zero-fitted coefficients.

From Figure 3.2, only limited predictors are selected for each liver, and the significant PC scores being selected in each liver do not show apparent similarity. One possible reason for this phenomenon can be the individual variation and the difference in preservation condition for each liver. Based on the predictor selection results of single livers, it is challenging to predict cross-liver quality by using only one logistic regression model. This is because the predictors selected in one liver may not be appropriate predictors for another. Therefore, the MTL approach [47] is introduced in the next subsection for cross-liver evaluation.

3.4.2. Cross-Liver Evaluation

3.4.2.1. Four Livers Overall Evaluation

The MR-MTL logistic regression model [70] was first applied to the overall four livers’ data, which include 4284 (1071 observations/liver × 4 livers) observations. Similar to the single liver modeling, 70% of the observations after PC scores extraction were randomly and stratified (equal number of training or testing observations selected from each liver) sampled from each liver as training data and the rest as testing data.
Table 3.1. Four livers overall MR-MTL logistic regression model classification error

<table>
<thead>
<tr>
<th></th>
<th>$\rho_1$</th>
<th>$\rho_2$</th>
<th>Type I (%)</th>
<th>Type II (%)</th>
<th>Overall Error (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training Data</td>
<td>9.6e-4</td>
<td>1.0e-5</td>
<td>0</td>
<td>4.84</td>
<td>3.47</td>
</tr>
<tr>
<td>Testing Data</td>
<td></td>
<td></td>
<td>0</td>
<td>4.71</td>
<td>3.49</td>
</tr>
</tbody>
</table>

Figure 3.3. Four livers overall MR-MTL logistic regression model goodness of fit. The dots are the predicated probabilities of liver viability. A liver is predicted as unviable with the probability of less than 0.5 and vice versa. The square marks are the unviable and viable liver observations with the probability labels of “0” and “1”.

Table 3.1 presents the training and testing errors of overall MR-MTL logistic regression model together with the tuning parameter selection results. With fairly small values of $\rho_1$ and $\rho_2$, both the training and testing errors are less than 4%, which indicates a strong correlation between the IR image and the liver quality. The MR-MTL logistic regression model is appropriate for modeling the cross-liver quality when all the livers are treated as an entirety. The small values of $\rho_1$ and $\rho_2$ describe the large discrepancy between livers and the great complexity of the model separately.

Figure 3.3 shows the goodness of fit of the MR-MTL model by comparing the calculated sample viability probability versus the actual viable or unviable results of the testing data. In accordance with the testing errors shown in Table 3.1, all of the viable liver observations are predicted correctly so Type I error is 0; and only a small portion of the unviable liver observations are
predicted as viable ones, so Type II error is as low as 4.71%. The Pearson residual analysis graph of the four livers overall MR-MTL model is shown in the Appendix B.

![Figure 3.4](image)

Figure 3.4. Four livers overall MR-MTL logistic regression model predictor selection. The significant PC scores in each liver following the order of PC score 1 to 4 in regions 1 to 6 from left to right, and liver #1 to #4 from top to bottom.

Figure 3.4 shows the predictor selection result of the PC scores of overall four livers, where the black-filled cells indicate the significant predictors to predict the liver quality by selecting the predictors of each liver that have the summation of coefficients greater or equal than 85% of the total summation of the coefficients. Although not all of the selected predictors are the same among four livers, 12 out of the 24 predictors are jointly selected in at least three livers, indicating certain degree of consistency of significant predictors for different livers, which also sets the basis of leave-one-liver-out modeling in the next subsection.

### 3.4.2.2. Leave-One-Out Evaluation

When a new liver becomes available, its time-dependent viability will be predicted by the available resources from exiting liver database. The leave-one-liver-out approach is applied in this section to imitate the potential liver evaluation process application by using three livers as training data and the forth one as testing data. The leave-one-liver-out evaluation was again conducted by MR-MTL [70]. Table 3.2 presents the four leave-one-liver-out results of tuning parameters selection, training error, and testing error. All of the training errors are still tightly controlled within 6%, while the testing errors show a wide range from 9% to 36%. This can be explained by the discrepancy among four livers and be validated by the fairly small values of $\rho_1$. 

34
Table 3.2. Leave-one-liver-out MR-MTL logistic regression model classification error

<table>
<thead>
<tr>
<th></th>
<th>$\rho_1$</th>
<th>$\rho_2$</th>
<th>Type I (%)</th>
<th>Type II (%)</th>
<th>Overall Error (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Leave-Liver 1-Out</td>
<td>2.1e-3</td>
<td>1.0e-5</td>
<td>Training</td>
<td>0</td>
<td>5.03</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Testing</td>
<td>16.27</td>
<td>6.19</td>
</tr>
<tr>
<td>Leave-Liver 2-Out</td>
<td>5.6e-5</td>
<td>1.0e-6</td>
<td>Training</td>
<td>0</td>
<td>7.43</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Testing</td>
<td>3.39</td>
<td>10.57</td>
</tr>
<tr>
<td>Leave-Liver 3-Out</td>
<td>1.6e-3</td>
<td>1.6e-5</td>
<td>Training</td>
<td>0</td>
<td>2.32</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Testing</td>
<td>1.36</td>
<td>48.84</td>
</tr>
<tr>
<td>Leave-Liver 4-Out</td>
<td>2.6e-3</td>
<td>1.0e-6</td>
<td>Training</td>
<td>0</td>
<td>5.46</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Testing</td>
<td>1.36</td>
<td>34.28</td>
</tr>
</tbody>
</table>

Figure 3.5. Leave-one-liver-out MR-MTL logistic regression model goodness of fit evaluation. (a) Leave-liver 1-out; (b) Leave-liver 2-out; (c) Leave-liver 3-out; (d) Leave-liver 4-out
Figure 3.5 shows the goodness of fit of each leave-one-liver-out model by plotting the calculated sample viability probability versus the actual sample viability for the testing livers. Unlike Figure 3.3, the $x$ axis direction here indicates the liver’s actual time-evolving process. The dots pattern in each liver shows the general pattern of the liver turning from viable to unviable. Except for the leave-liver 3-out graph (Figure 3.5 (c)), which shows an overall trend of the liver quality changing from viable to unviable, the rest three predication graphs all exhibited some discrepancies from the general pattern (Figure 3.5 (c)) at certain local sections during the observation time period. The Pearson residual analysis graphs of the leave-one-liver-out models are shown in the Appendix B.

Figure 3.6. Leave-one-liver-out MR-MTL logistic regression model predictor selection. (a) Leave-liver 1-out; (b) Leave-liver 2-out; (c) Leave-liver 3-out; (d) Leave-liver 4-out. The significant PC scores in each leave-one-out case following the order of PC scores 1 to 4 in Regions 1 to 6 from left to right, and livers #2, #3, #4 from top to bottom in (a), livers #1, #3, #4 from top to bottom in (b), livers #1, #2, #4 from top to bottom in (c), livers #1, #2, #3 from top to bottom in (d)
Figure 3.6 shows the predictor selection result of the PC scores of each leave-one-liver-out case, where the black-filled cells indicate the significant predictors to predict the liver quality by selecting the predictors of each liver that had the summation of coefficients greater or equal than 85% of the total summation of the coefficients. As shown in Figure 3.6, some similarities in predictor selection are observed among livers in modeling training. By evaluating the modeling results and predictor selection patterns, the MTL-based leave-one-liver-out modeling offers a reasonable testing error range with only four available liver data sets. Furthermore, this preliminary study points to a broad prospect of IR image application in the liver transplantation if more liver IR image data could be collected under controlled environment.

3.4.3. PC Loading Analysis

As the basic features extracted from IR images in this study, PC scores are significantly correlated with the liver viability. To explore the physical meaning implied in the PC score, PC loadings for each pixel of liver #1 are shown in Figures 3.7 for illustration. PC loadings for the rest of the livers are shown in the Appendix C. The color in each graph represents the magnitude of every pixel constituting each PC loading according to the color bar scale. Most figures show that the edges of the organ have higher PC loadings than the centered areas. This implies that the edges of the organ play a more important role to predict organ viability. This is consistent with the biomedical perception that liver deteriorates from the edge first [80]. Besides some edge areas showing significantly higher PC loadings, it is also noted that region 6 and the intersection of region 5, 3 and 1 illustrate evidently high PC loadings as well. We ascribe this phenomenon as the result of the function from a different organ or structure, which in this case region 6 is the bile area and the intersection of region 5, 3 and 1 is the fatty area. Since this paper only focuses on the relationship between liver viability and surface temperature, the conclusion of the effects of other
organ/structure will not be claimed here. Additional experiments will be needed to determine the hypothesis “bile and fatty region may have very strong impact on the viability of the liver” in the future research. By retrieving the IR images, one can not only model the quality change by the surface temperature of the organ, but also identify potential decayed regions based on the significance of the PC loadings in the organ.

Figure 3.7. Liver #1 PC loading graphs. (a) The 1\textsuperscript{st} PC loading; (b) the 2\textsuperscript{nd} PC loading; (c) the 3\textsuperscript{rd} PC loading; (d) the 4\textsuperscript{th} PC loading
3.5. Discussions

In this work, the IR image features are used as indicators to predict the liver viability within one liver and on new-coming livers. When the testing data and training data are both from one resource, i.e. the single liver evaluation in 3.4.1 and four livers overall evaluation in 3.4.2.1, the evaluation outcomes show accurate prediction results with testing errors all less than 5%. When the testing data and training data are from different resources like in the case of leave-one-liver-out evaluation in 3.4.2.2, testing errors have an acceptable but a wide range from 9 to 36%. We attribute the modeling performance difference to the liver-to-liver variation and probably the difference in preservation condition for each liver. The liver-to-liver variation can be verified in the single liver predictor selection results (Figure 3.2), where the important predictors selected from each liver do not show similar patterns. By adopting the MR-MTL method [70], the liver-to-liver variation is compromised by borrowing the commonality from each individual liver. Therefore, the four livers overall predictor selection results (Figure 3.4) show a certain level of consistency in terms of the number of joint-selected important predictors from four livers. Because of the co-existence of liver-to-liver variation and commonality, the leave-one-liver-out evaluation can be realized with varied but acceptable performances by the means of MR-MTL. In this case study, we used four livers to test the hypothesis. From the literature, the static cold storage is widely used to preserve organs since the beginning of organ preservation history, while machine perfusion with warm solution becomes a popular method in the past decades. It is still controversy to conclude the better method of the two for preservation [11, 81]. As a result, we did two experiments in the condition of 4°C, which is cold storage temperature, and did the other two in the condition of 22°C, which is room temperature. From Table 3.2 and Figure 3.5 in the paper, slight differences can be found between different temperature groups that the low temperature group has smaller prediction errors.
In addition, less variations are shown in low temperature group as well. However, the evidences are not strong enough to make the conclusion that the cold perfused method is better than the warm perfused method. More livers will be studied to determine the difference between cold perfused and warm perfused groups and to reduce the prediction error in future work. In particular, the donors’ health information will be used to quantify the heterogeneity of the organ sources, and reduce the variability in organ viability prediction.

The liver surface temperature and its distribution may not be the only determinant factors to affect the organ viability. The heterogeneity of porcine livers, perfusion conditions (e.g. perfusion temperature, perfusion solution, perfusion rate, etc.), preservation conditions (room temperature, humidity, etc.) are all possible variables need to be looked at in the future study. In this paper, we focus on the relationship between liver surface temperature indicated by IR images and liver viability during the perfusion stage. All the environmental factors like room temperature, humidity were strictly controlled. The experimental factors, except for the perfusion temperature, perfusion solution type and perfusion rate were also strictly controlled. The only fabricated confounding factor introduced in this case is the perfusion temperature, which was changed purposely to test the robustness of IR imaging-based model.

For future work, the online prediction can be investigated for providing real-time evaluation based on historical liver data (including both the training liver dataset and the up-to-time dataset of the current liver, which are used for model training) for liver transplantation. Other than liver quality evaluation, the IR image-based MTL logistic regression model can be applied to other type of organs, such as kidney to test the versatility of this proposed methodology. In addition, by developing this objective evaluation method, the biopsy process will be investigated to improve the quantifications of the viability of organs.
3.6. Conclusion

Organ transplantation is an important technique for prolonging lifetime when there are major organ failures. However, the preservation and evaluation of organ quality remain challenging issues. A non-invasive online organ temperature measuring system, along with computational thermographic organ quality evaluation models were evaluated in this work. This research has found that the liver surface temperature processed by PCA is a good predictor for evaluating the liver viability. Therefore, the IR imaging can be a very useful tool for liver quality evaluation. The single liver quality predictions have testing errors of zero by the logistic regression model with Lasso penalty, and the overall cross-liver quality prediction has a testing error as low as 5% with the application of the MR-MTL logistic regression model. Most importantly, the leave-one-liver-out predictions are able to predict the liver viability with testing errors varying from 9% to 36%, which show potential for improved liver quality evaluation.
Chapter 4. Modeling of Pre-transplant Liver Viability with Spatial-temporal Smooth Variable Selection

4.1. Introduction

Liver transplantation is the most effective medical solution for end-stage liver failure. Although there is a severe shortage of donated livers, liver transplantations have helped thousands of patients to prolong their life [82]. Although 158,588 patients received liver transplantations from 1988 to 2018, there are still 13,935 critically ill patients on the national Waiting List as of June 2018 [82]. It is extremely difficult to find a viable liver for transplantation, since the demand far exceeds supply; fewer than 20% of patients on the Waiting List actually receive one [82]. Many potentially transplantable livers are not used because of donor liver pathology (non-alcoholic steato-hepatitis, for example), the small window between procurement and implantation (<12 hours, generally) or imprecision of liver viability assessment methods [38].

Creating an accurate model for liver viability assessment is challenging but important, because an effective assessment method can increase the utilization of donors’ livers. Currently, liver assessment is primarily based on non-invasive visual inspections conducted by surgeons or invasive biopsy assessments conducted by pathologists. Visual inspection is subjective, and may easily yield inconsistency among experts [55], and the invasive biopsy is the most accurate method, yet may permanently damage liver tissues or cause liver dysfunction [5, 54]. The small, pre-transplantation biopsy may not accurately reflect the entire histomorphology of large organs such as liver, simply due to the size of the biopsy and whether the selected region for biopsy is representative of the entire organ. Therefore, an accurate non-invasive quantitative assessment method is needed.
By utilizing infrared (IR) imaging techniques to capture a series of liver images in a time order and using the pixel information from IR images as high-dimensional predictors, we developed and tested a spatial-temporal smooth variable selection (STSVS) method to model the correlation between IR images and liver viability. STSVS assesses the spatial-temporal correlation of IR image pixels to predict liver viability, and transforms the spatial-temporal correlation to the smoothness of model parameters over different locations on liver surface and over a time course. We believe, based on our studies, that STSVS performs better than benchmark computational models in predicting the liver viability due to its capacity of capturing both spatial and temporal effects that improves the overall model performance; current benchmark computational models do not consider effects that capture liver deterioration over time. Additionally, since STSVS is free of feature engineering that means this method does not need using domain knowledge to create features for modeling, the interpretable variable selection results could indicate the most important regions for predicting liver viability, potentially helping doctors and pathologists to understand the state of liver preservation/deterioration and define viable regions for split liver transplantation. Through the precise selection of transplantable livers potentially supported by this proposed method, more patients can benefit from the increased liver supply.

4.2. Background and Methods

4.2.1. Research Background
The effective use of a non-invasive, quantitative assessment method requires (1) a reliable and stable liver preservation procedure and (2) a technique capable of recording quantitative features from the preserved liver. To address both challenges, literature was reviewed on current methods for improving preservation effectiveness and providing non-invasive quantitative approaches in medical diagnosis [4, 83]. Machine perfusion systems (MPSs) have been developed for liver
preservation. In clinical use, these systems mimic some aspects of human physiology such as providing physiologically relevant waveforms and fluid circulation [43]. MPSs do not only reduce preservation injury but also provide a platform to evaluate the organ viability [84, 85]. Medical literature indicates that IR imaging is a potential technology that could be used for liver viability assessment, due to its non-invasive nature and its sensitivity to bodily dysfunction as a temperature indicator [57-59]. Gorbach et al. described the feasibility of evaluating porcine kidney quality by IR imaging under perfusion and applied IR imaging to correlate the surface temperature and the early critical renal ischemia during the pre-implantation stage [60, 61, 86]. However, image acquisition and processing was time-consuming and this was problematic.

With a large amount of thermographic data, there are many analytical approaches to assess liver viability described in the literature. As one of the most popular research topics in medical imaging interpretation, computer-aided diagnosis (CAD), provides means of addressing this task with both accuracy and consistency, because of its reliability in the capture of image features [14]. CAD reduces image ‘reading’ time [87], and is a major technique widely used in medical diagnosis and disease detection [15]. CAD has been applied for various medical imaging types, such as ultrasound, mammography, tomography, and magnetic resonance imaging to help with detecting breast lesions, breast cancer, lung nodules, and Alzheimer’s disease respectively [88-91]. However, the CAD-based machine learning approach needs a large sample size for accuracy; that is one of the major limitations preventing broad adaptation and use. In addition, the analytical results do not provide medical interpretation due to a lack of ranked, ‘important’ variable selections [16]. Statistical modeling has been widely used to evaluate medical tests for classification and prediction [18]. In organ transplantation, multivariate logistic regression was used to determine the decision-making involved with transplanting a split liver or a whole liver [62]. Cox regression modeling
has been used to investigate key factors related to liver failure in post-transplantation scenarios [63]. However, both methods focused on post-transplantation liver performance, rather than evaluating the liver viability before transplantation. In the pre-transplantation stage, people used Quantitative and Qualitative (QQ) models to predict liver viability by one continuous quantitative response representing the number of dead cells and one binary qualitative response representing the liver viability, simultaneously [5]. A multi-task learning logistic regression model was applied to evaluate cross-liver viability, using principal component analysis (PCA) to extract IR image features as predictors with binary responses representing viability [19]. The drawback of these methods is that they need complicated feature engineering to extract features from biopsy images or IR images, which is time-consuming. Moreover, the temporal effect representing liver viability changing over time has not been investigated yet.

Along the direction to use high-dimensional predictors and avoid the complex feature engineering in modeling, Tibshirani and co-workers reported a modification of the fused LASSO [92] method that provided a new perspective on modeling high-dimensional data without extensive feature engineering. Li et al. proposed a smooth spatial variable selection estimator, using spatial predictors from high-resolution images, to model the resistances of printed electronic circuits as indicators of print quality without feature extraction [93]. This method evaluated effects among neighborhood predictors on the quality response. However, due to limitations imposed by requirements of industrial quality inspection, a single microscopic image was captured only for the final product, so this method was limited to evaluation of spatial correlation. In order to address temporal effects that can be crucial in liver viability prediction, we further develop SSVS by using a series of IR images in a time order to capture the temporal effect, which is the STSVS method describe herein.
4.2.2. **Experiment Design**

Fresh porcine livers were obtained from a local abattoir and preserved on a custom-built MPS (CaVESWave®, BioMed Innovations LLC, Denver, NC). Figure 4.1 (a) shows the flow chart of a liver transplantation paradigm. Our work focuses on the perfusion and assessment procedures. As previously described, MPSs are reliable tools that can effectively maintain liver functions with minimal injury [94, 95]. In order to test the proposed STSVS method, porcine livers were used as substitutes of human livers and are a well-recognized translational model for transplantation studies. Porcine livers are similar to human livers in structure, size, physiology, and genomic constitution [65]. The livers were collected and put on ice within five minutes of death and thus chilled prior to return to the laboratory. Prior to transportation, livers were flushed with cold (4°C) phosphate-buffered physiologic saline using one-meter gravity infusion into the portal vein and hepatic artery. After arrival to the lab, livers were connected via hepatic artery and portal vein to the MPS as shown in Figure 4.1 (b). Figure 4.1 (c) shows the details about the machine perfusion system. Once the liver was connected to the MPS, phosphate-buffered physiologic saline was infused and a closed loop reperfusion system was established to mimic hepatic perfusion. Physiologically relevant systolic and diastolic pulse pressures, flows, and pulse rate were individually adjusted for the portal vein and hepatic artery. Livers were perfused for twenty-four hours and IR imaging data was collected for analysis. The experiment was repeated four times to collect the data of four different porcine livers (1, 2, 3, and 4). While the data from other sensors (e.g. organ’s pH value, systolic and diastolic blood pressure, turbidity, etc.) are collected, this paper mainly focuses on the correlation between the IR images and the liver viability, which is expected to not only predict the organ viability, but also reveal the local regions responsible for viability evaluation.
Static cold storage has been widely used to preserve livers since the beginning of liver transplantation. However, the optimal temperature for preservation is a matter of controversy. Some researchers have, however, argued that perfusion with warm solution is a more effective method [11, 81]. As a result, two livers were studied in cold perfusion (4°C) and two livers were studied in warm perfusion (22°C) in order to determine if perfusion temperature is a critical factor affecting liver viability during the preservation process. This indicates the importance to study the correlation of liver temperature and its viability.

Figure 4.1. Liver preservation. (a) flow chart of liver transplant, (b) procine liver, (c) machine perfusion system

47
4.2.3. Dataset

To collect IR images, a FLIR A655sc thermal IR camera (FLIR Systems Inc., Wilsonville, OR) was mounted above perfused porcine livers at a distance of one meter. For each porcine liver, surface temperature was captured as a \((640 \times 480)\) pixel image, as shown in Figure 4.2 by IR camera every 10 minutes during a 24-hour time period. 145 high-resolution \((640 \times 480)\) pixel images were collected as raw data assuming that each pixel in each image contains information of liver viability. Figure 4.2 the infrared images show the spatial-temporal change over time, and, in accordance with current liver preservation methods, we considered the 8-hour time point as the threshold of liver starting to turn from viable to unviable; this has been validated by studies of Perkins, et al. [5, 68].

In order to conduct more effective modeling, a data pooling [96] strategy was applied to combine nearby features into a small image that preserved the same important information as collected by measurements of single points alone, while reducing irrelevant details. Specifically, we calculated the average intensity of all pixels in each divided mesh from the original image and used it as the new mesh’s intensity [96]. Figure 4.3 shows that the pooled image of Liver 1 is reduced to a \((64 \times 48)\) pixel image as input data for modeling. The morphology of the image is preserved. Additionally, the pooled images of Liver 2, Liver 3, and Liver 4 are shown in Appendix D.

Figure 4.2. Image data
4.2.4. The Proposed Method

The assumptions of proposed methodology are: (1) a logistic regression model is adequate to model the relationship between the probability of the liver failure and spatial-temporal predictors assuming they are linear related. (2) The spatial-temporal correlation of all the pixels from infrared images can be transformed to spatial-temporal correlation of the model coefficients. (3) The model coefficients are associated with the size of grouped images accounting for the temporal effect, so the number of images included for model prediction has to be consistent with the size selected for temporal effect.

The STSVS method that we propose for this study is shown in Equation (4.1):

$$
\min_{\eta} \sum_{i=1}^{n} \left[ \log \left( 1 + e^{H_i^T \eta} \right) - z_i H_i^T \eta \right] + \rho_2 \| \eta \|_1 + \rho_3 \sum_{i=1}^{T} \sum_{o=1}^{M} \sum_{u=1}^{N} || S_{i,o,u} \circ \eta ||_1, \quad (4.1)
$$

where $H_i = (H_1, ..., H_J)^T$ representing image pixels, and $J$ is the number of predictors. $z_i$ is binary response that $z_i = 1$ represents a viable liver observation and $z_i = 0$ represents an unviable liver observation. $\eta = (\eta_{1,1,1}, ..., \eta_{T,M,N})^T = (\eta_1, ..., \eta_J)^T$ is the regression coefficient vector, $\rho_2$ is the tuning parameter to regularize the complexity of the model by shrinking the coefficients of less important predictors to zero, and $\|\cdot\|_1$ denotes the $\ell_1$-norm. $S_{i,o,u}$ is the three-dimensional smoothing tensor, each component of which corresponds to each predictor. The $S_{i,o,u} \circ \eta$ is the
inner product of the three-dimensional tensor $S_{i,o,u}$ with the vector $\eta$ (the summation of the element-wise product). The operation is conducted in such a way that the resulting scaler is calculated by taking the dot product of each component in $S_{i,o,u}$ with $\eta$. The first two terms in Equation (4.1) take the form of LASSO regression model, which is one of the most prevailing and efficient predictive model to address large dimensional predictors [69]. LASSO reduces the complexity of the regression model by penalizing the $\ell_1$-norm of model parameters. The logistic regression model with LASSO penalty $(\rho_2||\eta||_1)$ provides a good baseline of the prediction performance, but it lacks the ability to address the spatial-temporal effect. Since the liver viability deteriorates over time and the liver’s structure can be segregated into five lobes [71], it is rational to consider the correlation between predictors in a temporal order and group the spatially contiguous predictors together for having similar parameter values, the latter of which was first proposed in the work of Li et al. as SSVS method [93]. Therefore, the third term in Equation (4.1) is added to LASSO to take account of the spatial-temporal effect, and the spatial and temporal neighborhood correlation of each predictor is defined in Equations (4.2) to (4.4),

$$S_{i\neq l, w\neq o, v\neq u}^{i,o,u} = \begin{cases} 
0, & d(((t, w, v), (i, o, u)) > r \\
-\exp\left(-\frac{\sqrt{(t-i)^2+(w-o)^2+(v-u)^2}}{\sum_{h=1}^{r} \exp(-h)}\right), & d(((t, w, v), (i, o, u)) \leq r
\end{cases}, \quad (4.2)
$$

$$d(((t, w, v), (i, o, u)) = \sqrt{(t-i)^2+(w-o)^2+(v-u)^2}, \quad (4.3)
$$

$$S^{i,o,u} = -\text{sum}(S_{i\neq l, w\neq o, v\neq u}^{i,o,u}), \quad (4.4)
$$

where $i = 1, \ldots, T$, representing the number of images selected in a time order to reflect the temporal neighborhood effect (same for $t$); $T$ is chosen to be different values in computation to test the model performance; $o$ represents predictor (pixel) location in the horizontal direction of one image (same for $w$); $u$ represents the predictor (pixel) location in the vertical direction of one image (same for $v$); and $M$ and $N$ represent the total number of pixels in the horizontal and vertical
directions respectively. Therefore, Equation (4.3) calculates the Euclidean distance between two pixels at two time points. \((n \times T)\) equals the total number of images taken in the 24-hour period. \(r\) is a tuning factor, representing the neighborhood size selected to group together the correlated predictors. By incorporating the three-dimensional smoothing matrix, parameters of predictors are encouraged to be smoothing out in their spatial and temporal neighborhoods based on the criterion \((r\text{ value})\) set up for the model. \(\rho_3\) is the tuning parameter to regularize the parameter-smoothing term. The function of smoothing matrix is to assign a weighing factor for each predictor in a way that only the defined \(r\)-neighborhoods play effects to influence the estimation of the current predictor coefficient. By adding the spatial-temporal smooth term in addition to the LASSO penalty, the predictor coefficients tend to be similar to the values of its \(r\)-neighborhoods. SSVS is a special case of STSVS when \(T = 1\), so the three-dimensional smoothing tensor \(S_{i,0,u}\) becomes two-dimensional and the temporal effect is not considered [93]. Since the STSVS method adds on a temporal effect term to SSVS, it would outperform SSVS when predictors have certain inherent trend in the time domain with regard to responses.

To estimate the model parameters in Equation (4.1), the split Bregman method [97] is modified for the proposed estimator. Equation (4.1) can be expanded based on the augmented Lagrangian transformation, as

\[
\min_{\eta} \sum_{i=1}^{n} \left[ \log \left( 1 + e^{H_i^T \eta} \right) - z_i H_i^T \eta \right] + \rho_2 \|a\|_1 + \rho_3 \|b\|_1 + \langle g, \eta - a \rangle + \langle f, S\eta - b \rangle \\
+ \frac{\mu_1}{2} \|\eta - a\|_2^2 + \frac{\mu_2}{2} \|S\eta - b\|_2^2, \\
\text{Subject to } a = \eta, \\
\text{b} = S\eta, 
\]

where \(\langle \cdot, \cdot \rangle\) refers to the inner product of two vectors, \(g\) and \(f\) are dual variables corresponding to linear constraints of Equations (4.6) and (4.7); and \(\mu_1\) and \(\mu_2\) are parameters controlling the
convergence speed of the algorithm; $S$ is a matrix with each row being vectorized from $S^{i,o,u}$ for each variable parameter in $\eta$.

Solving the expanded objective function in Equations (4.5) to (4.7) equals solving Equation (4.8) by taking the two penalty terms out since they are functions of $\eta$ only.

$$\min_\eta \sum_{i=1}^{n} \left[ \log \left( 1 + e^{H_i^T \eta} \right) - z_i H_i^T \eta \right] + \langle g, \eta - a \rangle + \langle f, S\eta - b \rangle + \frac{\mu_1}{2} ||\eta - a||_2^2 + \frac{\mu_2}{2} ||S\eta - b||_2^2,$$

Equation (4.8)

The gradient descent method is used to solve Equation (4.8), shown as follows,

$$\eta^s = \eta^{s-1} - \delta_s \nabla f(\eta),$$

where the partial derivative of $\nabla f(\eta)$ with respect to $\eta$ is shown in Equation (4.10).

$$\nabla f(\eta) = \nabla \left( \sum_{i=1}^{n} \left[ \log \left( 1 + e^{H_i^T \eta} \right) - z_i H_i^T \eta \right] + \langle g, \eta - a \rangle + \langle f, S\eta - b \rangle + \frac{\mu_1}{2} ||\eta - a||_2^2 + \frac{\mu_2}{2} ||S\eta - b||_2^2 \right)$$

$$= \sum_{i=1}^{n} H_i \left( z_i - \frac{e^{H_i^T \eta}}{1 + e^{H_i^T \eta}} \right) + g + S^T g^{s-1} + \mu_1 (\eta - a) + \mu_2 S^T (S\eta - b).$$

Equation (4.10)

By substituting Equation (4.10) to Equation (4.9), $\eta$ is updated in each iteration until it is convergent in Equation (4.11). Equations (4.12) to (4.15) provide the formula of how $a, b, g, f$ are updated in each step by using soft threshold, where $Soft_{\rho_2}(\cdot)$ and $Soft_{\rho_3}(\cdot)$ are soft-threshold operators.

$$\eta^s = \eta^{s-1} - \delta_s \left[ \sum_{i=1}^{n} H_i \left( z_i - \frac{e^{H_i^T \eta^{s-1}}}{1 + e^{H_i^T \eta^{s-1}}} \right) + g^{s-1} + S^T f^{s-1} + \mu_1 (\eta^{s-1} a^{s-1}) + \mu_2 S^T (S\eta^{s-1} - b^{s-1}) \right]$$

Equation (4.11)

$$a^s = Soft_{\rho_2}(\eta^s + \frac{1}{\mu_1} g^{s-1})$$

Equation (4.12)

$$b^s = Soft_{\rho_3}(S\eta^s + \frac{1}{\mu_2} f^{s-1})$$

Equation (4.13)
\[ g^s = g^{s-1} + \mu_1(\eta^s - a^s) \]  
\[ f^s = f^{s-1} + \mu_2(S\eta^s - a^s) \]  

The extended Bayesian information criterion (EBIC) [98] is adopted to select the most appropriate tuning parameters of \( \rho_2 \) and \( \rho_3 \). The EBIC method is selected because its high accuracy when it is utilized in application dealing with high dimensional data [98].

4.3. Results and Discussions

We compared the STSVS prediction performance with results from the Generalized Linear Model (GLM) [99], Support Vector Machine (SVM) [100], LASSO [69], Fused LASSO [92] and SSVS [93] as benchmark models. GLM was selected as the most basic statistical modeling benchmark without any penalty terms in the model; LASSO was selected due to its capability of selecting significant predictors in a high dimensional dataset; fused LASSO further improved the variable selection performance of LASSO when predictors have an ordered pattern; SSVS considers the correlation between predictors in a spatial order and group the spatially contiguous predictors together for having similar parameter values; as the last benchmark model, SVM is a supervised learning model that typically used for classification. For STSVS, \( T \) is tuned from one to five to select the best value, which minimizes the prediction error. We randomly selected 70\% of the sample as the training dataset for model estimation and used the rest 30\% of the sample as testing dataset for model performance comparison. Additionally, 100 replications were performed for each method over each liver, and Figure 4.4 shows that the proposed STSVS method outperforms the benchmark models. The average prediction errors on the training and testing data for Livers 1 - 4 are summarized in Table E1 in Appendix E.
Figure 4.4. Prediction performance comparison

Only the best performance with the highlighted smallest testing error is presented for each liver. Testing error here is defined as the percentage of viable or unviable liver being predicted as unviable or viable among all observations in testing dataset, and training error is defined in a similar way as testing error but in training dataset. For Liver 1, Liver 2, and Liver 3, STSVS \((T = 2)\) has the lowest overall testing error among all methods, which is 7.43\%, 13.98\%, and 12.22\%, respectively. For liver 4, \(T = 3\) gives the best prediction performance which is 16.11\%. The difference optimal \(T\) value selected for Liver 4 compared with the rest livers could be due to the individual difference.

Figure 4.5 shows the variable selections of four benchmark models GLM, LASSO, Fused LASSO, and SSVS (STSVS with \(T = 1\)) for Liver 1. In 4.5 (a) and (b), the white pixels represent the selected significant variables from all \(64 \times 48\) predictors in a pooled IR image. In contrast, the black
pixels represent insignificant predictors to model performance. The color bar shows the value of significance of variables: the whiter the color the more significant the value. Variable selection results from GLM and LASSO are not useful in medical diagnostics based on image analysis because the selected pixels do not provide any graphical insights in terms of the shape of certain liver regions. Due to the non-interpretable nature, which is widely known as a “black box” strategy, SVM does not provide any medically relevant information [100]. Therefore, the variable selection image is not available for the SVM method.

Figure 4.5. Liver #1 variable selection results of (a) GLM, (b) LASSO, (c) Fused LASSO, (d) SSVS (STSVS with $T = 1$).
Figure 4.5 (c) and (d) show the variable selection results of Fused LASSO and SSVS (STSVS with $T = 1$) of Liver 1. The selected interpretable spatial variables for liver viability are medically meaningful because the shape of liver can be depicted and be distinguished from the background environment. Similar to the variable selection analysis of GLM and LASSO, the color represents the magnitude of beta coefficient estimates of each variable in a $64 \times 48$ pixel pooled IR image. The edges of the liver in variable selection figures have a brighter color than the central regions in a grayscale image, implying that the edges of liver play a more important role in predicting liver viability than central regions. This finding is consistent with the biomedical perception that liver easily deteriorates more rapidly from the edge regions than center regions [19, 80].

By conducting STSVS, the variable selection results can be further explained by considering the temporal effect. Figure 4.6 shows the magnitude of betas for STSVS with $T = 2$, meaning two consecutively taken IR images are simultaneously tied together to constitute a new observation data sample, and the color bar represents the value of magnitude of betas. This phenomenon represents the magnitude of variable coefficients changes over time, which could help define the deteriorating regions at different time. The selected model coefficients play a more significant role in determining the viability of the liver which can be seen from the increased brightness on the edges representing the problematic regions as the deteriorate level is worsen. Furthermore, the changes are consistent with medical domain experts’ claim that the deteriorating regions are at the edge locations to the adjacent regions gradually. In sum, STSVS does not only have the best prediction performance but also more interpretable. This proposed IR image-based liver viability evaluation method has the potential to be used in future studies to determine the mechanism of deterioration. The completed STSVS variable selection result for Liver 1 is shown in Appendix B.
Similar to the variable selection results obtained for Liver 1, Figure 4.7 shows the significant regions of Liver 2, Liver 3, and Liver 4 with $T = 2$ or $T = 3$. Findings are consistent with what are observed in Liver 1 that the variables around edges are more significant than central regions in determining the viability of livers. Although $T$ value offering the best prediction performance is not optimized to be the same for each liver, it never equals to one. The slightly different temporal effect on each liver’s viability could be related to individual liver’s features.
Figure 4.7. Variable selection results for (a) Liver # 2, (b) Liver #3 $T = 2$ (c) Liver #4 at $T = 3$

4.4. Conclusions

To save more severe liver disease patients, an effective liver assessment method is needed to increase the utilization of donor livers to benefit more patients and their families. A non-invasive IR imaging-based viability assessment method, along with STSVS, is proposed and validated in this paper. The spatial-temporal correlation shows that STSVS methodology provides the best viability prediction performance against benchmark models in our case study without feature
engineering to pre-process the dataset. The selected spatial-temporal features indicate the significant regions to predict liver viability.

For future work, the pre-existing medical conditions of donors will be investigated to model and address individual differences. The added individual features will help build a personalized model to further improve the prediction performance. IR imaging data could be fused with other online data to reveal a comprehensive picture of the targeting organ’s viability. Moreover, similarities among livers will be studied in both healthy regions and deteriorating regions for a better understanding of the liver deterioration pattern. In the long run, biopsy experiments will be performed to histologically verify the pathology present in deteriorating regions, and the latent relationship between surface temperature and internal temperature under skin will be determined to study the mechanism of deteriorating.
Chapter 5. Kidney Viability Forecasting by Multitask General Path Model in Transplantation

5.1. Introduction

Kidney transplantation is the best treatment option for patients who have kidney failure, because it can create the opportunity of living a longer and healthier life [101]. However, the demand for transplantable kidneys far exceeds the supply. Currently, <25% of total patients on the national kidney waiting list will ever receive a donated kidney; the median waiting time for a kidney is 3.6 years [102]. Even worse, the pool of kidney donors has deceased in recent years [103]. To match the donor kidney and recipient, United Network for Organ Sharing (UNOS) makes the decision based on three factors: distance, biological matching, and urgency [104]. However, the viability loss during transportation has never been considered as a key factor in decision-making and loss of viability may lead to >20% kidney discard rate [105]. Moreover, the lack of accurate methods to assess kidney viability causes geographic variation in kidney discard rate, with more discards in areas that have long distances between donated organs and potential recipients [106]. Therefore, a method of forecasting kidney viability is needed to support the decision-making to find the best recipient for kidney transplantation.

Forecasting kidney viability is very challenging for two reasons: 1) surgeons need to assess donor organ viability, but there are no universal guidelines or procedures for objectively and reliably assessing viability [107]. Currently, two methods, visual inspection and biopsy are used [5], and both are either subjective and/or invasive. Therefore, the assessment results are very likely to be inconsistent for different surgeons; 2) kidneys lose viability rapidly once they are procured, and the degradation rate is individual-specific [8]. Because assessment cannot be performed during
preservation or transportation, forecasting kidney viability needs to depend on historical data (i.e., how long can most kidneys remain viable in static cold storage, based on 1- and 5-year graft survival data). However, how to determine the information to be transferred from historical data of existing similar-but-non-identical kidneys to a new kidney is not a precise science and many donor factors and preservation conditions can affect viability data.

The objective of this research is to develop and test a method to forecast kidney viability during preservation by transferring knowledge from similar-but-non-identical kidneys to support decision-making of donor kidneys and recipients matching for predicting the best-fit patient. Moreover, the location and number of post-procurement biopsies can be optimized to reduce potential tissue damage due to repetitive biopsy procedures. Therefore, in this research, we formulate the viability forecasting problem as a multitask general path model (MT-GPM), where each task is one pre-defined region on the kidney, based on the vascular structure of the kidney, and a partial kidney can be transplanted for children [108]. This model incorporates advantages over multitask learning models and general path models by decomposing model coefficients into common effects and individual effects. Therefore, the heterogeneity across tasks can be captured to improve the forecasting accuracy while the commonality is derived from the aggregate of all tasks. More specifically, common effects of all similar-but-non-identical kidneys quantified by shared information are used to depict the trajectory of viability loss, and individual effects are used to identify the viability loss path difference of each kidney. The proposed method is validated by porcine kidney experiments using biopsy scores as the metric of the viability loss path. Compared to benchmark models including a multitask learning model (MTL) [109], a general path model (GPM) [110], and a general linear model (GLM) [99], MT-GPM has the best forecasting accuracy so that it can support decision-making.
The proposed MT-GPM can be used to accurately forecast kidney viability. Therefore, it can potentially help the kidney matching decision-making by selecting recipients that are more urgent but at greater distance, based on the estimation of viability loss. This method can be broadly applied to many domains as the response measurements can be expressed in a general function form, such as tissue deformation during needle biopsy. The generality is validated by a second case study in forecasting the magnitude of deformation path of needle biopsies used to support biopsy-based decision-making.

5.2. Research Background

There is a substantial body of research on kidney preservation and biopsy methods. Simple cold storage (CS), while widely used and effective, defines the preservation boundaries to generally less than 24 hours [111]. To prevent organ deterioration, the kidney preservation process has been improved by using machine perfusion systems (MPS), typically with flush solutions that buffers harsh molecular conditions that evolve during ischemia [112, 113]. On measure, MPS is the better solution for organ preservation because it significantly reduces the risk of delayed kidney function and potentially allows evaluating the viability of kidney before transplantation [4, 11]. Despite MPS offering a reliable platform for kidney preservation, in its current available iterations, it does not provide kidney viability measurements. To assess organ viability, biopsy is the most widely used method [114]. Biopsy is most accurate (predictive) when used with quantitative analysis of whole-side tissue images and sampling of several sites [115, 116]. Moreover, biopsy results are used extensively to predict organ quality for ‘non-standard’ donors; approximately 75% of extended criteria donor kidneys (ECD) are biopsied [117]. However, repeating biopsies may damage the kidney tissue permanently and this may potentially affect the utilization rate [118]. With the availability of MPS as platform and biopsy as validation method, non-invasive methods
can be used effectively to assess kidney viability. For example, computer-assisted diagnosis (CAD) methods can be applied to offer a quantitative diagnosis of the viability of the kidney. CAD has become one of the major research areas in medical diagnosis due to high speed, accuracy, and reliability [15]. Some researchers have used machine learning to predict the outcome of kidney transplantation, but the performance did not meet expectations [119]. Although machine learning is becoming increasingly sophisticated and more widely used, it requires a large sample size to keep training and modifying the model [16]. Because medical data is not always available, due to concerns for respecting the privacy of patients and the confidentiality of healthcare data [17, 120], statistical modeling is used to evaluate medical data for classification and prediction of small data sets [18].

To ensure reliability, data-driven degradation models are widely used to predict the useful lifetime of an engineering system or a biomedical system [110]. GPM, a degradation path model, was first proposed by Lu and Meeker [110, 121] for making inferences about the distribution of failure time for degradation data. One issue is that GPM is very sensitive to outliers. The unit-to-unit variability was modeled by adding random effects in the degradation model, and then developed with Bayesian updating to adjust the model performance [122, 123]. However, although the Markov Chain Monte Carlo (MCMC) procedure can be analytically used for intractable joint posterior density function estimation, it is computationally intensive for complex model structures. Zhou et al. introduced a degradation model in a nonparametric modeling framework, instead of using a common parametric model with random effects [124]. The limitation is that the non-parametric model is less efficient than the parametric model in some cases. Bagdonavicius et al. [125] investigated multiple failure modes in the degradation model. Hong et al. [126, 127] proposed dynamic covariates in the degradation model, using a linear random effects model, and then using
a nonlinear random effects model, but the general path model would only have good performance when the degradation curves are monotone or nearly monotone. GPM has been adopted from industrial reliability scenarios, assuming all the deployed equipment are identical in quality. However, the kidneys (biological ‘machines’) are heterogeneous due to size, shape, donor age, and molecular condition. Thus, GPM may not be applicable or appropriate to apply to kidney viability forecasting. Recently, a spatial-temporal degradation model was proposed to describe complex degradation items which show different degradation patterns at different locations [128, 129]. This model assumes the propagation of degradation at a constant speed. However, kidney heterogeneity may significantly affect outcomes from the use of this model.

Multi-task learning (MTL) is an approach for transferring information and improving generalization by using related tasks as an inductive bias [109]. The models define commonality simultaneously from all the tasks and exploit differences across tasks, so that acquired information can improve modeling performance [109]. MTL is widely used in many real-world applications, including medical and clinical fields. In organ transplantation, MTL logistic regression has been conducted to predict the liver viability based on infrared imaging data [19]. However, this method is not able to forecast the liver viability at a future time. Data Shared Lasso (DSL) [130] investigated the continuum between individual models for each group and one model for all groups, but DSL is not a forecasting solution and is not similar to MTL. In summary, the needs of kidney viability forecasting and the heterogeneity of different kidneys gave us the motivation to integrate GPM and MTL modeling methods to investigate the viability loss path. Thus, we propose MT-GPM as a means to forecast the viability of heterogeneous kidneys, in order to improve the kidney matching decision-making process.
5.3. Materials and Methods

In this research, we apply a data-driven model to forecast kidney viability, in terms of biopsy scores to support matching decision-making in order to reduce kidney discard rate. The assumptions of the model are: 1) the kidney viability loss path can be described in a general function form, for instance, the viability loss path in Case Study 1 can be approximated by quadratic functions based on historical data; 2) the kidney viability loss paths are similar-but-non-identical, because they share commonality among locations in different kidneys while maintaining individual differences.

5.3.1. The Proposed Method

We propose the MT-GPM that treat each location of four pre-defined locations on the kidney as one task. The biopsy score \( y \) of \( i \)-th viability loss path for task \( g \) at time \( t_j \) is given by:

\[
y_{(i,t,\pi_{(i,t)})} = \tau(\pi_j, \beta_{(i,t)}) + \varepsilon_{(i,t,\pi_{(i,t)})},
\]

where \( \pi_j \) is the time of the \( j \)-th biopsy score as measurement, \( j = 1, \ldots, n \); \( \varepsilon_{(i,t,\pi_{(i,t)})} \) is the measurement error which is assumed to independently and identically distribute as \( N(0, \sigma^2 \varepsilon) \); \( \tau_{(i,t)} \) is approximated quadratic path from assumption of the general form for \( i \)-th viability loss path in \( t \)-th task parameterized by \( \beta_{(i,t)}, i = 1, \ldots, I, t = 1, \ldots, G \). In each task, we have \( I \) viability loss paths representing the number of kidneys used in experiment. Because kidneys are similar-but-non-identical by sharing commonality and having heterogeneity. Therefore, we further decompose the model coefficient \( \beta_{(i,t)} \) to:

\[
\beta_{(i,t)} = \phi_{(t)} + \theta_{(i,t)},
\]

where \( \phi_{(t)} \in \mathbb{R}^{p(l+1)\times 1} \), \( p \) is the number of common-effects parameters representing commonality for all replicates; \( \theta_{(i,t)} \) is a vector of the \( i \)-th viability loss path individual-effects parameters in the \( t \)-th task representing heterogeneity. Thus, common-effects parameters are
transferring knowledge from all kidneys, and individual-effects parameters are identifying the kidney differences. Motivated by data shared lasso [130], our model is assumed as second order polynomial model with $\psi_{(i,t,\pi_{(i,t,j)},j)}$ representing timestamp of $i$-th viability loss path in $t$-th task as model inputs, so the $y_{(i,t,\pi_{(i,t,j)},j)}$ can be generated as:

$$y_{(i,t,\pi_{(i,t,j)},j)} = \psi_{(i,t,\pi_{(i,t,j)},j)}(\phi(t) + \theta_{(i,t)} + \varepsilon_{(i,t,\pi_{(i,t,j)},j)}).$$

The model parameters $\beta_{(i,t)} = \phi(t) + \theta_{(i,t)}$ can be estimated by optimizing Equation (5.4), and $r_t$ is given to control the amount of pooling, and we recommend $r_t = 1$ in this model:

$$\left(\hat{\phi}, \hat{\theta}\right) = \arg\min_{\phi, \theta} \frac{1}{2} \sum_{i,t,j} \left(y_{(i,t,\pi_{(i,t,j)},j)} - \psi_{(i,t,\pi_{(i,t,j)},j)}(\phi(t) + \theta_{(i,t)})\right)^2 + \rho_2 \left(\|\phi(t)\|_1 + \sum_{t=1}^T r_t \|\theta_{(i,t)}\|_1\right),$$

where the first part is the least square loss function to ensure the forecasting accuracy, and the second part is lasso penalty that penalizes on the sparsity of the coefficients [6]. Directly optimizing Equation (5.4) by using convex programming methods will leads to high computation workload and non-satisfactory time latency in forecasting due to multiple non-differentiable $l_1$ norms of model coefficients. Denoting $W_t = (\phi^T_t, \theta^T_1, \theta^T_2, ..., \theta^T_I) \in \mathbb{R}^{p(t+1) \times G}$ as the model coefficient vector, and $\varepsilon_{it}$ as the residual term, we show that optimizing Equation (5.4) can be approximated by optimizing Equation (5.6) by defining $X_t$ as:

$$X_t = \begin{pmatrix} \Psi_{1t} & r_1 \Psi_{1t} & 0 & \cdots & 0 \\ \Psi_{2t} & 0 & r_2 \Psi_{2t} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \Psi_{It} & 0 & 0 & \cdots & r_I \Psi_{It} \end{pmatrix},$$

where we separate the common effects and individual effects as the first column representing common effects and each $r_i \Psi_{it}$ is the covariates to estimate individual effects in one task among different viability loss paths.
\[
\hat{W} = \arg\min_W \sum_{t=1}^T \| X_t W_t - Y_t \|_F^2 + \rho_2 \| \tau \circ W \|_{2,1},
\]

(5.6)

where \( \tau \in \mathbb{R}^{p(t+1) \times G} \) is a known weights matrix, and the weights are data dependent \( \tau = 1/|\hat{\beta}_{OLS}|^\gamma \). If the weight is data-dependent, then the weighted lasso can have the oracle properties [131]. The \( l_{2,1} \)-norm regularized regression model is proposed to encourage multiple predictors to share similar sparsity patterns, which can quantify between-task similarities (i.e., common effects) and within task similarities (i.e., individual effects). The convexity of Equation (5.6) admits a globally optimal solution guaranteed by K.K.T. condition [132]; and \( \rho_2 \) is the tuning parameter to control the model sparsity, which leads to lower sample requirement [133]. The adaptive lasso achieves the oracle properties, and the shrinkage leads to a near-minimax-optimal estimator [131]. Therefore, it is applied to further improve the forecasting accuracy by adding weights to penalize different coefficients in the \( l_{2,1} \) penalty. The optimization problem showing in Equation (5.6) can be efficiently solved by accelerated gradient descent [134].

We use leave-one-kidney-out cross validation method to validate the model. Five kidneys out of six are used iteratively as training data and the remaining one as testing data. The initial model forecasts the viability by only using common effects, and then the model is updated by adding testing data one at a time to incorporate individual effects. The remaining test data is used to test forecasting accuracy.

In summary, we used a parametric model to represent the viability loss path. MT-GPM can forecast kidney viability in transplantation by simultaneously learning knowledge from all kidneys while identifying the individual differences. It can be extended to other domains, where the response measurements can be depicted by certain functions. Two case studies will be introduced to validate the generality of this method.
5.3.2. Case Study 1

To validate the proposed method, fresh porcine kidneys were procured. Pigs have many similar characteristics to humans in terms of heart rate, blood pressure, function and size, making use of the results from measurements of their kidneys readily translatable [135]. This study focused on the kidney viability assessment in preservation. Six porcine kidneys from young female pigs were collected from local abattoir, and were flushed with physiologic saline to remove residual blood. Next, the kidneys were transported back to the laboratory in a portable cooler, and they were immediately connected to a MPS [136] via renal artery and vein as below the Figure 5.1 (a) after arrival. Once the kidney was connected, more phosphate-buffered physiologic saline was infused and a closed-loop perfusion system was established to mimic renal blood perfusion. The perfusion time for each kidney was 12 hours in order to allow ample deterioration of the kidney over time, and a 12-hour time span is a similar comparison to the time transported from donor to recipient in reality. To accurately assess the viability of kidneys, each kidney was divided into four locations and biopsied as Figure 5.1 (b). We pre-defined each location as one task because of the vascular structure of the kidney. The locations closest to the artery and vein were perfused differently than the outer regions. Therefore, two locations, say, Locations 1 and 4 in Figure 5.1 (b) were at the extremities because they were at far-end of the artery and vein. In contrast, Locations 2 and 3 are more well-perfused regions.
The first biopsy was taken as soon as the kidney was connected to the machine, and then three more biopsies were procured with 4-hour interval at 4\textsuperscript{th} hour, 8\textsuperscript{th} hour, and 12\textsuperscript{th} hour. We intended to verify the viability loss during perfusion time but limit the number of biopsies, because repeating biopsies too many times at one location causes permanent damage to the kidney [116]. The time and location of biopsies for each kidney were randomized as Table 5.1 in order to reduce bias. Each biopsy was taken with a punch hole, and then was placed in 10\% neutral-buffered formalin to preserve the samples until evaluated histologically. The biopsy defect on the kidney was filled with surgical Gelfoam (Pfizer, Groton, CT) in order to prevent the damaging region expanding to other locations. All the biopsy samples were blindly graded by an experienced pathologist and ranked on a scale of 1, 2, and 3 representing bad, good, and excellent. Good and excellent biopsy scores signified the kidney was considered transplantable, otherwise they will be discarded. The biopsy scores were summarized in Table 5.1.
### Table 5.1. Kidney locations and biopsy scores summary

<table>
<thead>
<tr>
<th>Categories</th>
<th>0 Hour</th>
<th>4 Hour</th>
<th>8 Hour</th>
<th>12 Hour</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kidney 1</td>
<td>Locations</td>
<td>4&lt;sup&gt;th&lt;/sup&gt;</td>
<td>1&lt;sup&gt;st&lt;/sup&gt;</td>
<td>3&lt;sup&gt;rd&lt;/sup&gt;</td>
</tr>
<tr>
<td>Biopsy Scores</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Kidney 2</td>
<td>Locations</td>
<td>1&lt;sup&gt;st&lt;/sup&gt;</td>
<td>2&lt;sup&gt;nd&lt;/sup&gt;</td>
<td>3&lt;sup&gt;rd&lt;/sup&gt;</td>
</tr>
<tr>
<td>Biopsy Scores</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Kidney 3</td>
<td>Locations</td>
<td>1&lt;sup&gt;st&lt;/sup&gt;</td>
<td>4&lt;sup&gt;th&lt;/sup&gt;</td>
<td>2&lt;sup&gt;nd&lt;/sup&gt;</td>
</tr>
<tr>
<td>Biopsy Scores</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Kidney 4</td>
<td>Locations</td>
<td>4&lt;sup&gt;th&lt;/sup&gt;</td>
<td>1&lt;sup&gt;st&lt;/sup&gt;</td>
<td>2&lt;sup&gt;nd&lt;/sup&gt;</td>
</tr>
<tr>
<td>Biopsy Scores</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Kidney 5</td>
<td>Locations</td>
<td>1&lt;sup&gt;st&lt;/sup&gt;</td>
<td>4&lt;sup&gt;th&lt;/sup&gt;</td>
<td>3&lt;sup&gt;rd&lt;/sup&gt;</td>
</tr>
<tr>
<td>Biopsy Scores</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Kidney 6</td>
<td>Locations</td>
<td>2&lt;sup&gt;nd&lt;/sup&gt;</td>
<td>4&lt;sup&gt;th&lt;/sup&gt;</td>
<td>1&lt;sup&gt;st&lt;/sup&gt;</td>
</tr>
<tr>
<td>Biopsy Scores</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

Based on the information from above table, a 4 × 4 spatial-temporal biopsy score response surface representing biopsy scores of each kidney was generated. A monotonic Gaussian process model [137] was used to fit the data and generate the response surfaces because the deteriorating process is irreversible [138]. Figure 5.2 shows the six response surfaces with x, y, z-axis representing location, time and biopsy score for the six kidneys. Moreover, we assumed that the biopsy scores had spatial-temporal correlation. However, the sample size was limited. Therefore, interpolation was employed to increase the sample frequency from every 4 hours to every 30 minutes based on the response surfaces. As a result, six spatial-temporal biopsy score response surfaces were extended to 4 × 24 representing biopsies scores of each location in every 30 minutes.
5.3.3. Case Study 2

The proposed model is generalized to apply in other domains within a certain scope, in which a certain function form can roughly describe the response measurements. Needle biopsy, as a common medical tool, is applied as the assessment in transplantation to verify the viability of the kidney [4]. Some researchers investigated needle-tissue interaction mechanics to develop a theoretical model that quantifies the deformation of tissues and needles [139]. Since the magnitude of phantom deformation shows a path, the proposed method can be applied. To forecast the magnitude deformation path, the biopsy accuracy can be improved with online feedback guiding needle insertion process to obtain tissues from the target region. Moreover, this application can be potentially used in medical training and biopsy education.

The experiment setup shows as Figure 5.3 (a). The needle is placed above the phantom surface at 10 cm, and then inserted into the phantom. A polyvinyl alcohol hydrogel (PVA-H)-based phantom
was used to simulate the human tissue because its transparency made it easy to record the deformation of the tissue. The PVA-H phantoms were made of dimethyl sulfoxide and polyvinyl alcohol. The Young’s modulus of phantom was between 10Kpa and 50Kpa (similar to liver) and the Poisson ration is 0.49. A CCD camera was fixed in front of the phantom to capture markers inside the tissue, and 16 markers were made to record the displacements as Figure 5.3 (b). The photo sample frequency was two frames per second. In total, 26 photos were collected for analysis, and the displacements of x-axis and y-axis of each maker were calculated by imaging processing. The magnitude of marker deformation was used as the response in this case study. We treated each marker as one task, assuming each location has similar pattern, and four replicate measurements were made in each task.

![Figure 5.3. Needle insertion experiment](image)

(a) Experimental setup ![16 Markers Original Paths](image) (b) Magnitude of marker deformation

5.4. Results and Discussions

In Case Study 1, we applied the proposed method to analyze kidney viability data and compared forecasting accuracy with benchmark models including MTL, GPM, and GLM by root mean square error (RMSE), which quantifies the difference between forecasted biopsy scores and ground
truth. Because MTL, GMP, and GLM have benefits for regularization from cross-task effects, within-task effects, and general random effects, respectively. MT-GPM method decomposes common and individual effects representing cross-task and within-task effects in one modeling approach to improve forecasting accuracy.

The “leave-one-kidney-out” cross validation method is conducted to imitate the kidney viability forecasting by iteratively using five kidneys as training data and the remaining one as testing data. Therefore, in total, six scenarios are generated. MT-GPM can be applied without knowing any information of the new kidney by only using common effects from existing kidneys, and forecast the new kidney viability from the “cold start” stage. Each biopsy score is added to the training data following time sequence, and the model is updated by transferring information from new added biopsy scores. As more information about the new kidney is obtained, the model will perform better by considering individual effects. Therefore, the model forecasting will be performing more accurately as the RMSE continuously decreased. In Figure 5.4, we conclude that the proposed MT-GPM has the smallest overall prediction error compared to other benchmark models, and all six leave-one-kidney-out cross validation scenarios support the same conclusion. The RMSE are rapidly decreased to <0.3 in most of the six scenarios, because the common effects are so effective in model prediction that only few biopsy scores are needed to accurately estimate the model with individual effects. In Leave-1st –out (L1O), L4O and L5O scenarios, the increasing RMSEs of the first two or three timestamps show in Figure 5.4 is due to not having enough degrees of freedom for the quadratic model. Therefore, fluctuation exists when adding the first several biopsy scores.

In Figure 5.5 and Figure 5.6, we conclude that the model accuracy has increased as more biopsy scores from the testing kidney are added to the training data set. It proves that the model coefficients are updated as more information has been learned from the test kidney.
By applying MT-GPM model to kidney viability forecasting, the kidney matching decision can be improved by the estimation of kidney viability loss in preservation. Moreover, the best biopsy region of the kidney can be recommended for the pre-transplantation biopsy. For instance, the lowest ranked biopsy score region can be recommended to make sure the lowest biopsy score region is viable. Therefore, the entire kidney is transplantable. In contrast, the highest ranked biopsy score region can be recommended for biopsy if the overall viability is forecast to be bad. Thus, the practitioner can assure that the kidney should be discarded if the highest scored region is confirmed to be inviable after biopsy. As a result, the number of biopsies can be minimized to reduce the potential damaging effects of repeating biopsies.

In Figure 5.7, the graphs show the similarities among different tasks by comparing the Euclidean distances of model coefficients. In L1O, the task 1 is similar to task 4 which may represents location 1 and 4 have similar perfusion effect, and location 2 and 3 are similar due to the same reason. However, the similarity of locations can not be applied to the rest of kidneys due to heterogeneity issue. This finding has potential to be used in determining the perfusion effect with future validation experiment.
Figure 5.4. Forecasting accuracy (RMSE) of kidney viability case
Figure 5.5. “Cold start” predicted biopsy score and actual biopsy score comparison

Figure 5.6. Predicted biopsy score and actual biopsy score comparison at the last timestamp
MT-GPM has the best forecasting accuracy by distinguishing common effects and individual effects in one-step modeling approach, and it can potentially improve the effectiveness of kidney matching decision by forecasting the viability. In addition, the proposed model is generalized to not only benefit the kidney matching decision but also support other decision-making scenarios when the scope is such that a certain function form can roughly describe the response measurements.

Figure 5.7. Task similarities by Euclidean distance of model coefficients

In Case Study 2, Figure 5.3 (a) shows two groups of markers: red markers at top rows are close to the needle path having larger deformation and blue markers at bottom rows are further from the needle path with small deformation. Figure 5.3 (b) shows the true magnitude of maker deformation paths. Therefore, only mark 1 to 8 will be used for analytics since the magnitudes of deformation of mark 9 to 16 are too small. In addition, they cannot be represented by quadratic model. The model forecasting RMSEs are summarized in Figure 5.8. Similar to Case Study 1, the proposed MT-GPM outperforms the benchmark models in terms of forecasting accuracy. As the needle is continuously inserted into the phantom, more deformation path information was added to training.
data, so the model accuracy was significantly increased. Figure 5.8 shows that the proposed method in Leave-2nd-marker-out scenario has a different trajectory of deformation path compared to the rest, because the deformation path of marker 2 is significantly different from others showing in Figure 5.3 (b). However, MT-GPM can still address the heterogeneity issue and outperform the benchmark models. The proposed method can support needle biopsy decision-making, because it can learn the tissue deformation at different locations and interactive forces between the needle and tissue in order to provide meaningful and potentially real-time feedback for guiding needle insertion. The accuracy of pre-transplantation biopsy can be potentially increased to improve the overall transplantation performance. Model diagnostic and discussion of both case studies are in Appendix G.
Figure 5.8. Forecasting accuracy (RMSE) of needle biopsy case
5.5. Conclusions

In this research, we proposed the MT-GPM to forecast kidney viability during preservation in order to support kidney matching decision-making. The case study shows that the proposed method has satisfactory accuracy in forecasting kidney viability, and it outperforms the benchmark models including MTL, GPM, and GLM. The proposed method integrates MTL and GPM in a one-step modeling approach by learning the commonality from all tasks and capturing the heterogeneity by extrapolating each viability loss path. As a result, the forecasting performance can be improved. To learn the viability loss path of each kidney, it helps the kidney matching decision to be more efficient in providing customized kidney transportation services because the matching can be extended to facilitate greater distance but recognize the needs and potential success for recipients that are more urgent. In addition, this method can suggest the optimized location for conducting pre-transplantation biopsies and potentially saves more currently discarded kidneys for transplantation. The generality of the proposed method has been validated by the second case study to improve overall transplantation performance by providing online decision-making support in the pre-transplantation needle biopsy process.

Since MT-GPM is a new technique, follow-up studies need to be investigated in the future. For example, the kidney can be replaced with other organs, such as liver and lung, to test the universality of the proposed method. In addition, donor biological differences and medical records can be considered as covariates adding to the MT-GPM to further improve prediction performance.
Chapter 6. Conclusions and Future Research

The shortage of donor organs is a thorny issue that is yet to be solved. Because it is not realistic to increase the donor’s pool due to multiple reasons (e.g. legislation and religion issues), we need to increase the utilization of the available donated organs by improving the organ preservation and assessment. Therefore, the online data and information embedded in the preservation process need to be investigated. Due to the rapid development of sensing and information technology, biopsy images, infrared data and other types of in situ data were collected during the preservation process to reflect organ viability. Therefore, these available data provide a great opportunity to conduct data-driven modeling for organ viability prediction and forecasting to improve the accuracy and flexibility during the preservation process. However, these different types of data also bring challenges in data analytics of organ viability. The lack of non-invasive and objective methods to assess viability of heterogeneous organs gives the motivation for this study. In this dissertation, several data-driven predictive and forecasting models are investigated to improve organ viability assessment. In particular, the following methods are proposed:

1) Quantitative and qualitative models is proposed to provide an objective assessment method for organ viability based on biopsy image features, which has validated the biopsy results.

2) A multitask learning logistic regression model is proposed to predict the liver viability based on spatial infrared data of liver surface. This work identifies that the liver surface temperature is a good predictor for assessing the liver viability.

3) A spatial-temporal smooth variable selection is proposed to address the spatial-temporal correlations in the infrared data of liver surface without feature engineering. The selected spatial-temporal features indicate the significant regions on the liver to predict the liver viability.
4) A multitask general path model is proposed to forecast kidney viability during preservation to support kidney matching decision-making. It helps the kidney matching decision to be more efficient because the matching can be extended to facilitate greater distance but more urgent recipients. The generality of this method is validated by needle biopsy case study.

The proposed methods provide an applicable solution to evaluate the organ viability during preservation process to reduce the organ discard rate. First, the proposed methods play as supplements to the traditional biopsy method to quantitatively validate and interpret the biopsy results. Second, the infrared camera is used to collect surface temperature data of the organs to predict viability for the first time in literature. The non-invasive nature of infrared camera provides the ideal solution to evaluate the organ viability without damaging organs. Third, the proposed data-driven models can quantify the correlation between organ viability and predicting variables to evaluate the organs. Fourth, the organ forecasting method significantly pushes the boundaries of current evaluation methods by learning the viability loss to improve the organ matching decision-making. Moreover, the pre-transplantation biopsy location and time can be recommended by the proposed method to minimize the potential damages on organs. Thus, this closed-loop organ viability prediction and forecasting method is completed to improve the overall performance in pre-transplantation stage from organ procurement to preservation, and it is convenient to implement. Therefore, this systematic solution can reduce the organ discard rate and potentially help more end-stage patients to regain health.

There are several future research topics can be studied to further improve the overall organ transplantation performance, and the graphical statement is in Appendix H.

1) Based on the assessment methods in this dissertation, a research regarding online organ preservation condition optimization can be conducted to control the organ preservation
environment. The preservation environment is critical to preserve graft function to improve long-term transplantation outcome, however the best preservation environment setting, such as preservation temperature, type of solution, and the selection of preservation machine, is still not clear [140]. Therefore, experiments of different preservation conditions can be applied to preserve a certain organ, and different settings will be recorded accordingly. The mitochondrial function of the organ can be used to provide a sensitive analysis of cellular metabolic function to compare different preservation environments [141]. By modeling the mitochondrial function and environment setting variables, the preservation condition could be optimized to extend the effectiveness on preserving the organ metabolic function.

2) To improve the modeling performance of proposed methods, the pre-existing medical conditions of donors will be investigated to address individual differences. Therefore, the new personalized model will be used to further improve prediction performance, which is also a popular research topic in medical field [142, 143]. In additional, the significant regions retrieved from Chapter 4 will be used to investigate the latent relationship between the surface temperature and internal temperature under skin to further study the mechanism of organ deteriorating.

3) By applying the proposed methods to estimate the organ viability, Figure 6.1 shows that the preservation decision-making regarding preservation methods based on distance and patient urgency can be added to the current UNOS’s decision-making flow to improve the overall decision-making process and its outcomes. This multistage decision-making can potentially optimize the organ transplantation decision by applying decision analysis techniques including risk assessment, multi-objective trade-off assessment, and preference elicitation [144,145,146].
Figure 6.1. The proposed decision-making flow for UNOS
References


Chen J, Chen Z. Extended BIC for small-n-large-p sparse GLM. *Statistica Sinica* 2012; 555-574.


Appendixes

Appendix A. Liver Quality Evaluation Results based on The Full Data Set (1431 Observations)

Table A1 and Figure A1 show the single liver evaluation results. Table A2 and Figure A2 show the four livers overall evaluation results. Table A3, Figure A3 and A4 show the leave-one-liver-out evaluation results. Results based on the full data set are in line with the results from the truncated date set shown in Section 3.4.1 and 3.4.2 but with inferior performance.

Table A1. Single liver logistic regression model classification error

<table>
<thead>
<tr>
<th>Liver #</th>
<th>Training Type I (%)</th>
<th>Training Type II (%)</th>
<th>Testing Overall Error (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Liver #1</td>
<td>0.29</td>
<td>0</td>
<td>0.10</td>
</tr>
<tr>
<td></td>
<td>0.76</td>
<td>0</td>
<td>0.23</td>
</tr>
<tr>
<td>Liver #2</td>
<td>0</td>
<td>0.15</td>
<td>0.10</td>
</tr>
<tr>
<td></td>
<td>0.72</td>
<td>0</td>
<td>0.23</td>
</tr>
<tr>
<td>Liver #3</td>
<td>0.87</td>
<td>0.15</td>
<td>0.40</td>
</tr>
<tr>
<td></td>
<td>0.76</td>
<td>0.67</td>
<td>0.70</td>
</tr>
<tr>
<td>Liver #4</td>
<td>0</td>
<td>0.15</td>
<td>0.10</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0.67</td>
<td>0.47</td>
</tr>
</tbody>
</table>
Figure A1. Pearson residual analysis of each liver in logistic regression models. (a) Liver #1; (b) Liver #2; (c) Liver #3; (d) Liver #4

Table A2. Four livers overall MR-MTL logistic regression model classification error

<table>
<thead>
<tr>
<th></th>
<th>$\rho_1$</th>
<th>$\rho_2$</th>
<th>Type I (%)</th>
<th>Type II (%)</th>
<th>Overall Error (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training Data</td>
<td>1600</td>
<td>1.0e-5</td>
<td>0</td>
<td>26.95</td>
<td>18.03</td>
</tr>
<tr>
<td>Testing Data</td>
<td></td>
<td></td>
<td>0.18</td>
<td>27.34</td>
<td>18.26</td>
</tr>
</tbody>
</table>
Figure A2. Residual Analysis. (a) Four livers overall MR-MTL logistic regression model residual analysis; (b) Four livers overall MR-MTL logistic regression model goodness of fit evaluation

Table A3. Leave-one-liver-out MR-MTL logistic regression model classification error

<table>
<thead>
<tr>
<th></th>
<th>ρ1</th>
<th>ρ2</th>
<th>Type I (%)</th>
<th>Type II (%)</th>
<th>Overall Error (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Leave-Liver 1-Out</td>
<td>2100</td>
<td>1.0e-5</td>
<td>Training</td>
<td>0</td>
<td>20.05</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Testing</td>
<td>41.68</td>
<td>46.23</td>
</tr>
<tr>
<td>Leave-Liver 2-Out</td>
<td>1600</td>
<td>1.0e-5</td>
<td>Training</td>
<td>0</td>
<td>27.86</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Testing</td>
<td>4.84</td>
<td>32.11</td>
</tr>
<tr>
<td>Leave-Liver 3-Out</td>
<td>2100</td>
<td>1.0e-5</td>
<td>Training</td>
<td>0.07</td>
<td>20.12</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Testing</td>
<td>2.11</td>
<td>71.44</td>
</tr>
<tr>
<td>Leave-Liver 4-Out</td>
<td>1600</td>
<td>1.0e-5</td>
<td>Training</td>
<td>0</td>
<td>28.03</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Testing</td>
<td>0</td>
<td>49.16</td>
</tr>
</tbody>
</table>
Figure A3. Leave-one-liver-out MR-MTL logistic regression model residual analysis. (a) Leave-liver 1-out; (b) Leave-liver 2-out; (c) Leave-liver 3-out; (d) Leave-liver 4-out
Figure A4. Leave-one-liver-out MR-MTL logistic regression model goodness of fit evaluation. (a) Leave-liver 1-out; (b) Leave-liver 2-out; (c) Leave-liver 3-out; (d) Leave-liver 4-out
Table A4. Classification errors comparison between the full data set and the truncated data set

<table>
<thead>
<tr>
<th>Time Period</th>
<th>Observations</th>
<th>Type I (%)</th>
<th>Type II (%)</th>
<th>Overall Error (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Leave-Liver 1-Out</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0-5 hour</td>
<td>1071</td>
<td>16.27</td>
<td>N/A</td>
<td>16.27</td>
</tr>
<tr>
<td></td>
<td>1431</td>
<td>15.25</td>
<td>N/A</td>
<td>15.25</td>
</tr>
<tr>
<td>6-11 hour</td>
<td>1071</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>1431</td>
<td>85.00</td>
<td>0</td>
<td>42.74</td>
</tr>
<tr>
<td>12-24 hour</td>
<td>1071</td>
<td>N/A</td>
<td>6.19</td>
<td>6.19</td>
</tr>
<tr>
<td></td>
<td>1431</td>
<td>N/A</td>
<td>56.81</td>
<td>56.81</td>
</tr>
<tr>
<td><strong>Leave-Liver 2-Out</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0-5 hour</td>
<td>1071</td>
<td>3.39</td>
<td>N/A</td>
<td>3.39</td>
</tr>
<tr>
<td></td>
<td>1431</td>
<td>7.80</td>
<td>N/A</td>
<td>7.80</td>
</tr>
<tr>
<td>6-11 hour</td>
<td>1071</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>1431</td>
<td>0</td>
<td>100</td>
<td>49.72</td>
</tr>
<tr>
<td>12-24 hour</td>
<td>1071</td>
<td>N/A</td>
<td>10.57</td>
<td>10.57</td>
</tr>
<tr>
<td></td>
<td>1431</td>
<td>N/A</td>
<td>16.58</td>
<td>16.58</td>
</tr>
<tr>
<td><strong>Leave-Liver 3-Out</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0-5 hour</td>
<td>1071</td>
<td>1.36</td>
<td>N/A</td>
<td>1.36</td>
</tr>
<tr>
<td></td>
<td>1431</td>
<td>0</td>
<td>N/A</td>
<td>0</td>
</tr>
<tr>
<td>6-11 hour</td>
<td>1071</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>1431</td>
<td>5.56</td>
<td>97.75</td>
<td>51.40</td>
</tr>
<tr>
<td>12-24 hour</td>
<td>1071</td>
<td>N/A</td>
<td>48.84</td>
<td>48.84</td>
</tr>
<tr>
<td></td>
<td>1431</td>
<td>N/A</td>
<td>65.42</td>
<td>65.42</td>
</tr>
<tr>
<td><strong>Leave-Liver 4-Out</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0-5 hour</td>
<td>1071</td>
<td>1.36</td>
<td>N/A</td>
<td>1.36</td>
</tr>
<tr>
<td></td>
<td>1431</td>
<td>0</td>
<td>N/A</td>
<td>0</td>
</tr>
<tr>
<td>6-11 hour</td>
<td>1071</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>1431</td>
<td>0</td>
<td>53.93</td>
<td>26.82</td>
</tr>
<tr>
<td>12-24 hour</td>
<td>1071</td>
<td>N/A</td>
<td>34.28</td>
<td>34.28</td>
</tr>
<tr>
<td></td>
<td>1431</td>
<td>N/A</td>
<td>48.07</td>
<td>48.07</td>
</tr>
</tbody>
</table>
Table A4 shows the testing errors differences between the full data set (1431 observations) vs. the truncated data set (1071 observations) modeling results. The “N/A” cells in Table A4 refer to the “not applicable” errors, indicating one or more particular types of error do not exist during the assigned period of times. It is evidently observed that the testing errors based on the full data set during the truncated time period of 6-11 hour due to the uncertainty of the liver quality are significantly higher than the beginning (0-5 hour) and ending (12-24 hour) time period. The comparison results rationalize the exclusion of data points in 6-11 hour for the modeling work shown in this study.
Appendix B. Pearson Residual Analysis of Single Liver and Cross-Liver Evaluation Models

Figure B1 shows the single liver Pearson residual analysis by plotting the Pearson residual in \( y \) axis against training sample observations in \( x \) axis for each liver to validate the model. All of the residual plots show random patterns, indicating that the logistic regression model with Lasso penalty perform well for single liver evaluation.

![Liver #1 Pearson Residual Analysis](image1)

![Liver #2 Pearson Residual Analysis](image2)

![Liver #3 Pearson Residual Analysis](image3)

![Liver #4 Pearson Residual Analysis](image4)

Figure B1. Pearson residual analysis of each liver in logistic regression models. (a) Liver #1; (b) Liver #2; (c) Liver #3; (d) Liver #4

Figure B2 validates the four livers overall MR-MTL model by plotting the Pearson residual against the training sample observations. In Figure B2, similar pattern of residuals around zero followed
by a period of negative residuals appear four times in the plot, which represents a similar residual trend for each liver in the overall MR-MTL modeling.

Figure B2. Four livers overall MR-MTL logistic regression model residual analysis

Figure B3 validates each leave-one-liver-out model. Similar to Figure B2, every Pearson residual graph shows three repeated patterns for the data sets from three training livers. Those small portions with Pearson residuals around zero are in accordance with the 0% Type I errors; while the curve-shaped portions represent fairly large training errors which are reflected in the Type II errors (shown in Table 3.2).
Figure B3. Leave-one-liver-out MR-MTL logistic regression model residual analysis. (a) Leave-liver 1-out (L-1-O); (b) Leave-liver 2-out (L-2-O); (c) Leave-liver 3-out (L-3-O); (d) Leave-liver 4-out (L-4-O)
Appendix C. PC Loading Graphs for Livers #2, #3, and #4

The following figures (Figure C1 to C3) show similar patterns as in Figure 3.7, implying that the edges of organ with higher PC loadings start deteriorating first.

Figure C1. Liver #2 PC loading graphs. (a) The 1st PC loading; (b) the 2nd PC loading; (c) the 3rd PC loading; (d) the 4th PC loading
Figure C2. Liver #3 PC loading graphs. (a) The 1st PC loading; (b) the 2nd PC loading; (c) the 3rd PC loading; (d) the 4th PC loading.
Figure C3. Liver #4 PC loading graphs. (a) The 1st PC loading; (b) the 2nd PC loading; (c) the 3rd PC loading; (d) the 4th PC loading
Appendix D. Image Pooling Results for Liver #2, #3, and #4

Figure D1 shows the image pooling results for liver #2, #3, and #4.

(a)

(b)

(c)

Figure D1. Image pooling results for (a) liver #2 (b) liver #3 (c) liver #4
### Appendix E. Numerical Prediction Performance

Table E1 shows the prediction performance of all four kidneys by STSVS method.

**Table E1. Prediction performance of Liver #1, Liver #2, Liver #3, and Liver #4**

<table>
<thead>
<tr>
<th>Methods</th>
<th>Liver #1</th>
<th></th>
<th>Liver #2</th>
<th></th>
<th>Liver #3</th>
<th></th>
<th>Liver #4</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>GLM</td>
<td>0.00%</td>
<td>12.78%</td>
<td>13.24%</td>
<td>24.33%</td>
<td>12.21%</td>
<td>19.44%</td>
<td>10.49%</td>
<td>21.21%</td>
</tr>
<tr>
<td>SVM</td>
<td>1.50%</td>
<td>10.11%</td>
<td>14.35%</td>
<td>20.17%</td>
<td>9.32%</td>
<td>15.31%</td>
<td>8.98%</td>
<td>19.83%</td>
</tr>
<tr>
<td>LASSO</td>
<td>0.40%</td>
<td>10.67%</td>
<td>12.51%</td>
<td>19.23%</td>
<td>9.98%</td>
<td>16.21%</td>
<td>9.53%</td>
<td>17.36%</td>
</tr>
<tr>
<td>Fused LASSO</td>
<td>5.95%</td>
<td>8.33%</td>
<td>12.27%</td>
<td>19.78%</td>
<td>9.69%</td>
<td>15.34%</td>
<td>9.11%</td>
<td>17.11%</td>
</tr>
<tr>
<td>STSVS ($T=2$)</td>
<td>1.99%</td>
<td><strong>7.43%</strong></td>
<td>9.23%</td>
<td><strong>13.98%</strong></td>
<td>8.21%</td>
<td><strong>12.22%</strong></td>
<td>9.69%</td>
<td>18.21%</td>
</tr>
<tr>
<td>STSVS ($T=3$)</td>
<td>1.20%</td>
<td>8.69%</td>
<td>10.70%</td>
<td>17.56%</td>
<td>8.45%</td>
<td>14.23%</td>
<td>9.08%</td>
<td><strong>16.11%</strong></td>
</tr>
</tbody>
</table>
Appendix F. Liver #1 STSVS Variable Selection Results

Figure F.1 shows the variable selection results for STSVS with \( T = 2, 3, 4, 5 \) accordingly, meaning two, three, four, and five IR images are simultaneously put together to constitute a new observation data sample. As \( T \) exceeds 2, the shade patterns do not change significantly any more compared with the second one, as shown in Figure F.1 (b) to (d).

![Variable selection results for STSVS](image)

(a) \( T=2 \), (b) \( T=3 \), (c) \( T=4 \), and (d) \( T=5 \)

Figure F1. Liver #1 variable selection results for STSVS at (a) \( T=2 \), (b) \( T=3 \), (c) \( T=4 \), and (d) \( T=5 \)
Appendix G. Model Diagnostic and Discussions

Figure G1, each $6 \times 6$ matrix shows pairwise dissimilarities among model coefficients estimated from six kidneys, where each row/each column represents model coefficients estimated from one of the six kidneys. The off-diagonal entry at the $i$-th row and the $j$-th column represents the model coefficient similarity between $i$-th and the $j$-th kidneys, $i = 1, \ldots, 6, j = 1, \ldots, 6$. In Figure G1, darker blue representing greater dissimilarity between kidneys. In contrast, lighter blue means more similarity in model coefficients. By visualizing this connectivity matrix, the dissimilarities among different kidneys can be determined based on color intensity. Therefore, it can easily conclude that the third kidney is significantly different from the rest of the kidneys.

![Figure G1. Pairwise distance between model coefficients in Case Study 1](image-url)
In each leave-one-kidney-out scenario, each $4 \times 4$ matrix in Figure G2 to Figure G7 shows pairwise dissimilarities among model coefficients estimated from four tasks. The heterogeneity of each location can be indicated by the model coefficients from the heat map.

Figure G2. L1O pairwise distance between model coefficients in Case Study 1
Figure G3. L2O pairwise distance between model coefficients in Case Study 1

Figure G4. L3O pairwise distance between model coefficients in Case Study 1
Figure G5. L4O pairwise distance between model coefficients in Case Study 1

Figure G6. L5O pairwise distance between model coefficients in Case Study 1
Similar to Case Study 1, each $8 \times 8$ matrix in Figure G8 shows pairwise dissimilarity among model coefficients estimated from eight markers in Case Study 2. It can conclude that the model of the second marker shares the least commonality in model coefficients with the rest of markers, which is in coincidence with L2O model performance comparison in Figure 5.5. As discussed in previous section, the trajectory of the second marker deformation path is significantly different from the other markers.

Figure G7. L6O pairwise distance between model coefficients in Case Study 1
As discussed in the main body of this paper, only mark 1 to 8 will be used for analytics since the magnitudes of deformation of mark 9 to 16 are too small. In addition, they cannot be represented by quadratic model. Figure G9 demonstrates the difference of two groups by their model coefficients.
Figure G9. Pairwise distance between model coefficients in Case Study 2 (16 markers)
Appendix H. Future Research Discussions

Figure H9. Future research discussions