Automated Decision Support System for Traumatic Injuries

by

Negar Farzaneh

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in The University of Michigan
2021

Doctoral Committee:

Professor Kayvan Najarian, Chair
Professor Brian Athey
Associate Professor Alan Boyle
Professor Harm Derksen
Associate Professor Maureen Sartor
Assistant Professor Craig A. Williamson

Negar Farzaneh

negarf@umich.edu

ORCID iD: 0000-0003-1200-5274

Dedicated to my parents, Mitra and Khodarahm.

ACKNOWLEDGEMENTS

I am incredibly grateful to the people who made this dissertation possible through their mentorship, support, friendship, and love. Primarily, I would like to thank Dr. Kayvan Najarian for giving me an academic home as a PhD student. I am so deeply grateful for his mentorship, support, empathy, and for challenging me to push past my limits. Thank you for your role in my formation as a scholar and as a person.

Thank you to the rest of my thesis committee members for their invaluable input and support throughout this process: Dr. Brian Athey, Dr. Alan Boyle, Dr. Harm Derksen, Dr. Maureen Sartor, and Dr. Craig. Williamson. I would like to particularly thank Dr. Williamson for being incredibly generous with his time and advice every time I sought him out. Many of the ideas in this dissertation stem from my conversations with him. I am grateful to Dr. Athey for all his insightful comments and sometimes tough questions. Specifically, his questions during my first dissertation committee meeting challenged me to always consider the innovation, contribution, and clinical relevance of my research. Ever since, I have asked myself the same questions at every turn of my graduate studies as they help me to better understand how my research fits into the bigger picture.

My sincere thanks go Dr. Reza Soroushmehr for setting me on this path from the beginning of my graduate studies and accompanying me the entire way. Thank you for your patience, guidance, and support throughout this research process, and thank you for always seeing the best in everyone — including me. I would also like to acknowledge the crucial role of Dr. Jonathan Gryak for his help and attention to all the little details that have gone into making my doctoral research possible. I truly appreciate his time, knowledge, and coordination efforts.

I have had the privilege of working with Dr. Erica Stein. I am extremely grateful and indebted to her for her expertise accompanied by her sincere and valuable guidance that were essential in determining the direction of the abdominal trauma assessment component of my research.

I wish to acknowledge the support that I received from the Department of Computational Medicine and Bioinformatics administrative staff and faculty. Special thanks to Dr. Margit Burmeister, Helen Severino, and Julia Eussen. Dr. Burmeister was readily available and eager to address my many questions and concerns. Thank you for your generous and honest advice during my doctoral study. I am grateful to Helen Severino for helping me to navigate the complicated process of submitting and receiving my NIH F31 application. She made the process seamless. Thank you, Julia Eussen, for her first-class support and willingness to handle any issues along the way.

I thank my fellow labmates: Cheng Jiang, Narathip Reamaroon, Mohsen Hooshmand, Elyas Sabeti, Heming Yao, Carrie Li, Sardar Ansari, Larry Hernandez, Lu Lu, Alexander Wood, and Craig Biwer for their friendship, stimulating discussions, sleepless nights working together before deadlines, and all the fun we have had in the past six years. A big thank you to Vy Nguyen for her support and companionship from the moment I joined the program. She helped me to adjust to American culture and has always been willing to read and give feedback on pages of my writing, which also includes this very acknowledgement section.

I am so proud of all the accomplishments of the UM DCMB Girls Who Code Club in the past four years, which was possible due to the contribution of every single member. I am inspired by Brooke Wolford, who shaped the way I approach outreach and education by introducing me to the UM DCMB Girls Who Code Club. Her enthusiasm, hard work, and grit shaped the team and made the club a great success.

I am most grateful to my parents for their love, guidance, and sacrifices to educate and prepare me for my future. Over the past six years, when I was apart from them, my parents spent countless hours with me on the phone, even if we were each silently doing our own work. Just knowing that they were on the other side of the line made me feel at home, safe, and focused. I wish to thank my grandmother for being an inspirational source of strength and wisdom in the family. Finally, I must express my love and appreciation to Arya Farahi, my best friend and my husband, for his unwavering support and encouragement.

TABLE OF CONTENTS

CHAPTER

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF APPENDICES

Appendix

ABSTRACT

With trauma being one of the leading causes of death in the U.S., automated decision support systems that can accurately detect traumatic injuries and predict their outcomes are crucial for preventing secondary injuries and guiding care management. My dissertation research incorporates machine learning and image processing techniques to extract knowledge from structured (e.g., electronic health records) and unstructured (e.g., computed tomography images) data to generate real-time, robust, quantitative trauma diagnosis and prognosis. This work addresses two challenges: 1) incorporating clinical domain knowledge into deep convolutional neural networks using classical image processing techniques and 2) using post-hoc explainers to align black box predictive machine learning models with clinical domain knowledge. Addressing these challenges is necessary for developing trustworthy clinical decision-support systems that can be generalized across the healthcare system. Motivated by this goal, we introduce an explainable and expert-guided machine learning framework to predict the outcome of traumatic brain injury. We also propose image processing approaches to automatically assess trauma from computed tomography scans.

This research comprises four projects.

In the first project, we propose an explainable hierarchical machine learning framework to predict the long-term functional outcome of traumatic brain injury using information available in electronic health records. This information includes demographic data, baseline features, radiology reports, laboratory values, injury severity scores, and medical history. To build such a framework, we peer inside the black-box machine learning models to explain their rationale for each predicted risk score. Accordingly, additional layers of statistical inference and human expert validation are added to the model, which ensures the predicted risk score's trustworthiness. We demonstrate that imposing statistical and domain knowledge "checks and balances" not only does not adversely affect the performance of the machine learning classifier but also makes it more reliable.

In the second project, we introduce a framework for detecting and assessing the severity of brain subdural hematomas. First, the hematoma is segmented using a combination of hand-crafted and deep learning features. Next, we calculate the volume of the injured region to quantitatively assess its severity. We show that the combination of classical image processing and deep learning can outperform deep-learning-only methods to achieve improved average performance and robustness.

In the third project, we develop a framework to identify and assess liver trauma by calculating the percentage of the liver parenchyma disrupted by trauma. First, liver parenchyma and trauma masks are segmented by employing a deep learning backbone. Next, these segmented regions are refined with respect to the domain knowledge about the location and intensity distribution of liver trauma. This framework accurately estimated the severity of liver parenchyma trauma.

In the final project, we propose a kidney segmentation method for patients with blunt abdominal trauma. This model incorporates machine learning and active contour modeling to generate kidney masks on abdominal CT images. The resultant of this component can provide a region of interest for screening kidney traumas in future studies.

Together, the four projects discussed in this thesis contribute to diagnosis and prognosis of trauma across multiple body regions. They provide a quantitative assessment of traumas that is a more accurate measurement of the risk for adverse health outcomes as an alternative to current qualitative and sometimes subjective current clinical practice.

CHAPTER I

Introduction

Trauma is the primary cause of mortality for individuals younger than 46 years old and the leading cause of years of life lost in the U.S. (*Rhee et al.*, 2014). Traumatic Brain Injury (TBI) and abdominal trauma contribute to most of this traumatic mortality and morbidity (*Maas et al.*, 2008; *Faul et al.*, 2010; *Jansen et al.*, 2008; *Taghavi and Askari*, 2019).

A traumatic brain injury occurs when direct or indirect trauma disrupts the brain's normal function. TBI which contribute substantially to traumatic mortality and morbidity, are frequently referred to as the silent epidemic (*Faul et al.*, 2010). Each year, about 1.7 million people sustain TBIs in the U.S., of whom approximately 52,000 die and another 126,000 experience long-term impairment or disability (*Faul et al.*, 2010). One-third of patients who died because of TBI were able to talk or obey commands before their death. This suggests the effect of secondary injuries, rather than the initial injuries at the time of trauma, is an important contributor to mortality (*Moppett*, 2007). Thus, the early diagnosis, prognosis, and management of TBI, particularly during the "golden hour", the period of time following the incident, could significantly minimize mortality and progression of secondary injuries. TBI can cause Subdural Hematoma (SDH), diffuse axonal injury, cerebral contusions, lacerations, and other injuries. SDH is one of the most common sequelae of TBI and the most frequent indication for craniotomy following trauma. Hence, its early diagnosis and evacuation can improve patient outcomes. In addition to early detection of TBI, early and accurate prediction of TBI long-term functional outcomes is important for guiding physicians and families through care management and early resuscitation.

Traumatic abdominal injury can lead to multiple complications including laceration of major organs such as the liver and kidneys. Due to its anterior location, size, and fragile parenchyma, the liver is the most frequently injured abdominal organ involved in blunt abdominal trauma (*Ahmed and Vernick*, 2011; *Taghavi and Askari*, 2019; *Badger et al.*, 2009; *Arumugam et al.*, 2015). In fact, approximately 5% of all trauma admissions are attributed to liver trauma (*Taghavi and Askari*, 2019). However, clinical signs alone are insufficient to diagnose stable abdominal injuries, and missed injuries are a common cause of morbidity and late mortality (*Jansen et al.*, 2008).

One of the most important initial tests performed on trauma patients to evaluate injury severity is Computed Tomography (CT) imaging (*Wintermark et al.*, 2015; *Badger et al.*, 2009; *Buquicchio et al.*, 2018). It provides critical information such as severe brain or liver hematomas that are found to be significant predictors of trauma severity and outcome (*Wintermark et al.*, 2015; *Badger et al.*, 2009; *Buquicchio et al.*, 2018). Their existence may specify the urgent need for surgical intervention, which if delayed may lead to death. Traditional visual examinations of CT scans are still the only investigation done by clinicians, which are time consuming, prone to human error, and costly. Moreover, traumatic injury severity is conventionally assessed by measurements derived primarily from a single 2D CT slice. For example, the maximum depth of a hemorrhagic region in one slice can be used to determine the severity of liver laceration, or the thickness of a lesion will be used to assess brain hemorrhage. This kind of measurement, while simple, may not exploit other characteristics of the injured area, such as volume, which could potentially make diagnosis more accurate.

Based on this information, there is a need for an automated, generalizable decision support system to improve diagnosis accuracy and prevent delayed or misdiagnosis by extracting knowledge from heavily underused data sources. My dissertation aims to fill this need by developing algorithmic solutions for a trauma decision-support system. This system has two general components: outcome prediction and image processing.

1.1 Outcome Prediction

During the outcome prediction phase, patient-level data from Electronic Health Records (EHR) was used to develop predictive models to support and enhance clinical decision-making for TBI.

Prognosis of the long-term functional outcome of TBI is essential for personalized management of that injury. Nonetheless, accurate prediction remains unavailable. Although machine learning has shown promise in many fields, including medical diagnosis and prognosis, such models are rarely deployed in real-world settings due to lack of transparency and trustworthiness. To address these drawbacks, we sought to develop a machine learning-based framework that is explainable and aligns with clinical domain knowledge.

1.2 Image Processing

During the image processing phase, we used CT scans to detect and quantify two key abnormalities in the brain and abdominal cavity regions: brain subdural hematoma and liver parenchymal disruption. In addition, kidney masks are segmented as a region of interest for screening for kidney traumas. Current image processing methods can be categorized into two main groups: 1) classical image processing techniques mainly driven by clinical domain knowledge about the underlying disease and 2) deep learning techniques that are end-to-end black box models. Deep convolutional neural networks have recently been successfully applied to medical image segmentation tasks and have been proven to outperform classical models in terms of average performance (*Esteva et al.*, 2021). However, in failure cases, these deep learning models may make mistakes that classical models will not. This non-robustness can be attributed to lack of reasoning and might primarily affect the minority cases (*Esteva et al.*, 2021) because of automation bias. In this thesis, we focused on integrating clinical domain knowledge into the deep learning model. The resulting algorithms can act as safeguards or triage tool for alerting radiologists and clinicians to potentially moderate and severe injuries.

1.3 Specific Aims

In Chapter II, we propose a novel framework to predict the long-term functional outcome of TBI patients. This framework incorporated additional layers of statistical inference and human expert validation to create an intelligible machine learning framework. To build our intelligible model, we explained, validated, and accordingly revised a machine learning classifier to align it with clinical domain knowledge. To explain the rationale behind the decision-making process of the machine learning model, we used SHAP (SHapley Additive exPlanations) post-hoc explainer that estimates the contribution of each variable to the final decision at both patient and population levels. Next, we validated the contribution of each variable at the population level. We identified the variables that were non-robust or showed counterintuitive behaviors and excluded them from the variable set. Contents of this chapter are accepted for publication (*Farzaneh et al.*, 2021b).

In Chapter III, we investigate the use of deep learning to diagnose and assess SDH for patients with potential TBI. We first segmented SDH from CT images. This automated segmentation of SDH was then used to quantify the injury using a volumetric measurement, which enabled the detection of moderate and severe SDH subjects. For the SDH segmentation task, we developed a new combinatory pipeline that benefits from both classical image processing and deep learning approaches. The proposed algorithm employs a joint feature representation of domain knowledge-driven hand-crafted features and data-driven deep learning features to train a random forest model that achieved greater accuracy and robustness compared to deep learning-only methods. The hand-crafted features in this algorithm reflected human domain knowledge that can compensate for the limitations in deep model performance on unobserved regions of the input data distribution, leading to this improved result. Contents of this chapter were partially published in (*Farzaneh et al.*, 2017b) and (*Farzaneh et al.*, 2020).

Chapter IV moves on from TBI and focuses on developing an automated framework to identify and assess liver trauma from contrast-enhanced CT scans. This framework started by segmenting initial masks of both liver parenchyma (including both normal and affected portions) and regions affected by trauma. Next, during the post-processing step, we integrated human domain knowledge about the location and intensity distribution of liver trauma into the model to avoid false positive regions. After generating the liver parenchyma and trauma masks, liver parenchymal disruption involvement was computed as the volume of liver parenchyma that was disrupted by trauma. The liver parenchymal disruption involvement is one of the main criteria in determining the American Association for the Surgery of Trauma (AAST) liver injury scale - the primary tool currently in use to assess the extent of the liver trauma and guide management. Contents of this chapter were partially published in (*Farzaneh et al.*, 2017a) or are under review for publication in (*Farzaneh et al.*, 2021a).

In Chapter V, we focus on developing a kidney segmentation method for trauma injury patients. We first used machine learning classifiers to detect an initialization mask inside each kidney. The boundary of each mask was then evolved using an active contour modeling technique. The results indicates an accurate performance of the proposed model in segmenting kidneys in CT scans with various contrast phases. The resultant of this project provides the region of interest to search for kidney traumas. Contents of this chapter were partially published in (*Farzaneh et al.*, 2016) and (*Farzaneh et al.*, 2018).

Finally, Chapter VI provides an overall discussion of the results of this dissertation, potential impacts of the developed models, and possible future directions.

CHAPTER II

Predicting the Long-Term Functional Outcome of Traumatic Brain Injury Patients

2.1 Introduction

Traumatic Brain Injury (TBI), often referred to as the "silent epidemic", is the leading cause of death among young Americans (*Faul et al.*, 2010; *Rhee et al.*, 2014). While accurate early prognostication of TBI outcomes can guide physicians and families through early resuscitation and treatment planning, such a prognostic system remains unavailable. After initial resuscitation, most patients with severe brain injury die as a result of withdrawal of life-sustaining treatment. Consequently, there is a critical need for accurate tools that can identify and prevent early withdrawal from treatment in severe TBI patients who still have a reasonable chance for a favorable outcome (*Hemphill III and White*, 2009; *Geurts et al.*, 2014). Even experienced neurosurgeons and neurocritical care practitioners frequently overestimate the likelihood of poor neurological outcome in comparison with validated prediction scores (*Moore et al.*, 2013). By accurately predicting long-term functional outcomes, physicians can make more evidence-based and informed decisions in such cases.

Over the last four decades, several studies aimed to produce prognostic models by using patient responsiveness (*Committee on Medical Aspects of Automotive Safety*, 1971; *Teasdale and Jennett*, 1974; *Wijdicks et al.*, 2005), radiographic images (*Maas et al.*, 2005; *Marshall et al.*, 1991; *Stenberg et al.*, 2017), or said images in combination with other risk factors (*Stenberg et al.*, 2017; *Steyerberg et al.*, 2008; *Collaborators*, 2008; *Junior et al.*, 2017; *Rizoli et al.*, 2016; *Hukkelhoven et al.*, 2005). However, the accuracy and applicability of these models over complex and heterogeneous cohorts are questionable (*Deepika and Shukla*, 2016; *Majdan et al.*, 2017). A primary reason for the failure of these models is the oversimplification of the risk assessment method employed, in which only a limited number risk factors are considered.

Artificial intelligence has shown great promise in enhancing the medical decision-making process, specifically when there is a significant complexity and uncertainty involved with the risk assessment task (*Rajkomar et al.*, 2019). Machine learning algorithms enable integrating multiple sources of information in a complex non-linear fashion for accurate data-informed prognostication. A few recent studies sought to tackle the oversimplification in previous TBI prognosis studies by employing machine learning methods (*Rau et al.*, 2018; *Matsuo et al.*, 2020). However, this approach comes with a trade-off: a sophisticated machine learning model's rationale for an individual decision is not readily interpretable by clinicians. The black box nature of such algorithms prevents them from being integrated into medical practice where transparency is imperative(*Elshawi et al.*, 2019; *Fogel and Kvedar*, 2018; *Vellido*, 2019; *Kelly et al.*, 2019; *Holzinger et al.*, 2017). Acceptance of such models by clinicians in real-world settings requires the underlying reasoning of a model to be explainable, understandable, and trustworthy (*Kelly et al.*, 2019; *Holzinger et al.*, 2017; *Caruana et al.*, 2015).

Another concern is the susceptibility of machine learning models to poor performance over unobserved data, which is also known as a major problem called overfitting. This is particularly acute in medical applications where, due to privacy and intellectual property issues, it is costly and often impractical to have an ideal data set that is sufficiently large and heterogeneous to represent all subtypes of the condition under study. Thus, during the training stage, machine learning can potentially learn unrealistic cohort-specific patterns that are not generalizable (*Kelly et al.*, 2019) or clinically significant. Such models introduce additional sources of bias to the prediction model (*Kelly et al.*, 2019), which will not be readily detectable if employed in a black box fashion.

In this chapter, we propose a novel machine learning framework that incorporates additional layers of statistical inference and human expert validation to create an *intelligible model* for predicting long-term functional outcomes of TBI patients using data available at the time of hospital admission. Inspired by Caruana et al. (*Caruana et al.*, 2015), an intelligible model is defined as a model that is both interpretable and aligned with clinical domain knowledge. The proposed machine learning framework constructs an intelligible model through performance explanation, human expert validation, and final model training. To explain the decision-making process of the machine learning model, Shapley values were used to estimate the contribution of each variable to the final decision (*Lundberg and Lee*, 2017; *Lundberg et al.*, 2018). Next, the contribution of each variable was clinically validated at the population level, with variables determined to be non-robust or exhibiting counterintuitive behaviors subsequently excluded. The results of this process suggest that including counterintuitive features introduces bias to the model. To further explore this hypothesis, a case study was performed on one of the features with counterintuitive behavior in which its impact on model bias was analyzed.

2.2 Materials

In this study, as in most recent TBI clinical trials, the long-term functional outcome after TBI is assessed using the Glasgow Outcome Scale-Extended (GOSE), a global scale for functional outcomes, at 6 months after injury. The original Glasgow Outcome Scale (GOS) and its more detailed and recent revision, the GOSE, are the most widely accepted systems to rate TBI outcomes, having been used in more than 90 percent of high-quality TBI randomized trials. The GOSE has been extensively validated, is the most widely cited measure of acute brain injury outcomes, and is recommended by both the US National Institutes of Health (NIH) and the UK Department of Health (*McMillan et al.*, 2016). In this study, GOSE 1–4 (death, persistent vegetative state, and severe disability) were regarded as unfavorable outcomes, while GOSE 5-8 (moderate disability, and good recovery) correspond to patients with favorable outcomes.

This is a secondary analysis of the Progesterone for the Treatment of Traumatic Brain Injury III (ProTECT) data set that includes adults who experienced a moderate to severe brain injury caused by blunt trauma (*Wright et al.*, 2014). This multi-site, randomized clinical trial (ClinicalTrials.gov identifier NCT00822900) ensured that all racial, ethnic, and geographical variations were appropriately incorporated. Patients were excluded from ProTECT III if they had an initial Glasgow Comma Scale (GCS) of 3, bilateral dilated unresponsive pupils, or were otherwise determined to have non-survivable injuries. The data set includes electronic data for 882 patient (*Wright et al.*, 2014). Among the 882 patients, 831 met the inclusion criteria. Of 831 individuals admitted to the hospital, 348 were identified to have experienced poor outcomes, with the remaining 483 attaining a favorable recovery at six months.

A rich source of patient-level information is available in the EHR contained within the ProTECT III data set. This information includes demographic data, baseline features, radiology reports, laboratory values, injury severity scores, and medical history. Demographics and clinical characteristics of the patient cohort are summarized in Table 2.1. In this study, only data available at the time of hospital admission was used.

Table 2.1: Demographic and clinical characteristics of study subjects.

Characteristic	**GOSE $\leq$ 4**	**GOSE >4**	***p*-value**
Total Subjects, n	348	483	
Female, n	97 (27.87%)	127 (26.29%)	>0.05
Age, median [Q1, Q3]	45 [29, 59]	31 [22, 45]	<0.001
Abbreviated injury score [Q1, Q3]	29 [22, 36]	22 [14, 29]	<0.001
Head injury severity Score, median [Q1, Q3]	4 [4, 5]	3 [3, 4]	<0.001
Cause of injury			
Motor vehicle collision, n	99 (28.45%)	204 (42.24%)	<0.001
Motorcycle/scooter/ATV/bicycle crash, n	82 (23.56%)	126 (26.09%)	>0.05
Pedestrian struck by moving vehicle, n	66 (18.97%)	42 (8.70%)	<0.001
Fall, n	63 (18.10%)	70 (14.49%)	>0.05
Assault, n	24 (6.90%)	22 (4.55%)	>0.05
Other or Unknown, n	14 (4.02%)	19 (3.93%)	>0.05
Initial Glasgow Coma Scale			
Motor response, median [Q1, Q3]	4 [3, 5]	5 [4, 5]	<0.001
Eye opening response, median [Q1, Q3]	1 [1, 2]	2 [1, 3]	<0.001
Verbal response, median [Q1, Q3]	1 [1, 2]	2 [1, 2]	<0.001

2.3 Methods

2.3.1 Outcome Prediction Framework and Experimental Design

The proposed framework for TBI outcome prediction is outlined in Fig. 2.1. First, a machine learning classifier is trained to predict the risk score for each patient (Section 2.3.3), with SHAP (SHapley Additive exPlanations) values being used to explain the model's predictions (Section 2.3.4). Next, the global behavior of the SHAP values for each input variable is evaluated, with only those shown to be statistically robust selected for further consideration. These robust variables are then validated by a multidisciplinary team of clinical experts, with features with counterintuitive behavior investigated further and excluded to avoid potential sources of bias (Section 2.3.5.2). Although excluding counterintuitive variables might negatively affect the overall accuracy, it is an essential step to develop a sensible and trustworthy algorithm that can be used operationally in a clinical setting.

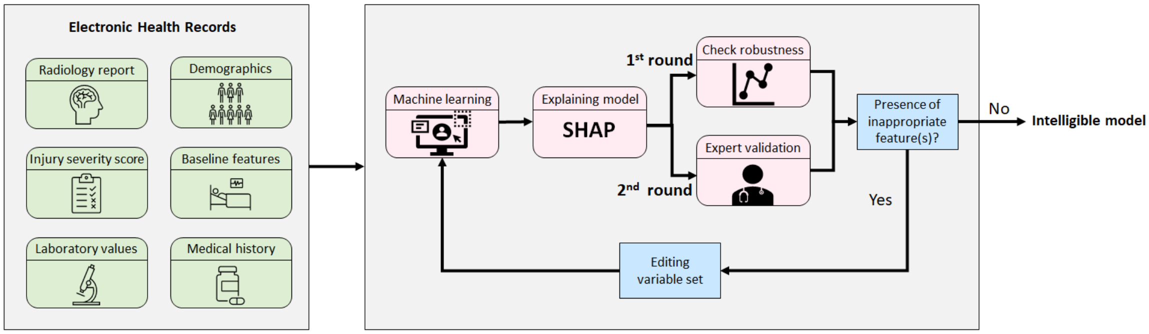


Figure 2.1: The proposed framework for developing an intelligible TBI prognostic model. After training an initial machine learning model, input features are selected based on statistical robustness and clinical validity. The machine learning model is retrained after each step of feature selection (*Farzaneh et al.*, 2021b).

The experimental design is outlined in Fig. 2.2. To ensure that the final prognostic model is not affected by the test set, 25% of the whole data set will be randomly selected and set aside. This test data set will remain untouched until the prognostic model is trained and finalized.

2.3.2 Data Pre-processing

First, among all EHR variables available in the ProTECT III data set, those available at the time of admission were selected. Next, these selected variables were reviewed by an expert board-certified physician in neurology and neurocritical care, with all those deemed clinically relevant chosen as input variables. Information regarding race or ethnicity was excluded, as this information could reinforce an unwanted retrospective bias and/or discrimination rather than a direct cause of the predicted outcome (*Paulus and Kent*, 2017, 2020).

Missing values were replaced by the average of available cases in the training set.

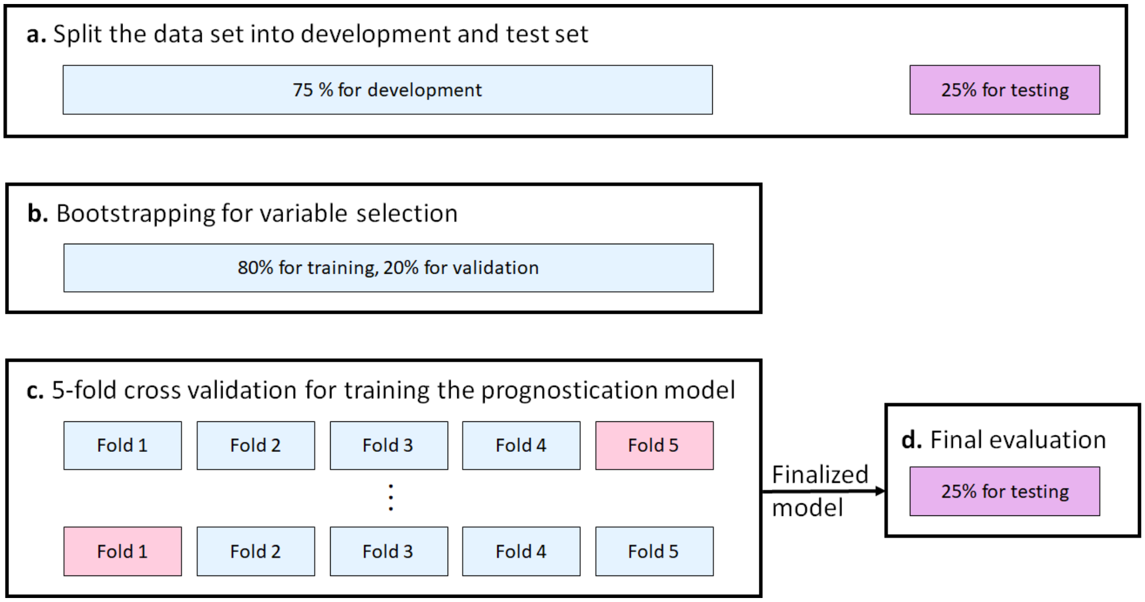

a. Split the data set into development and test set
75 % for development
25% for testing
b. Bootstrapping for variable selection
80% for training, 20% for validation
c. 5-fold cross validation for training the prognostication model
Fold 1
Fold 2
Fold 3
Fold 4
Fold 5
Fold 1
Fold 2
Fold 3
Fold 4
Fold 5
Finalized model
d. Final evaluation
25% for testing

Figure 2.2: Experimental design and evaluation strategy for developing an intelligible TBI prognostic model. (a) At study onset, 25% of the data was set aside for final evaluation. The remaining 75% was used to develop the model, either in (b) the bootstrap step for variable selection, or (c) training the prognostic model using 5 fold cross-validation (*Farzaneh et al.*, 2021b).

2.3.3 Machine Learning Module

The XGBoost (eXtreme Gradient Boosting) algorithm was employed to classify each patient as experiencing either a favorable or unfavorable outcome at 6 months and to estimate its corresponding probability. More information regarding the selection of the XGBoost algorithm is available in Appendix A and Table A.2. XGBoost is a sequential tree growing algorithm with weighted samples. Compared to other boosting methods, XGBoost incorporates regularization parameters, making it relatively robust against noise and outliers while reducing over-fitting. The XGBoost package in Python was used to build this prognostic model (*Chen and Guestrin*, 2016).

The hyperparameters were optimized for all models using grid search over a specified subset of the hyperparameter space. For each model, the combination of the hyperparameters that yielded the maximum F1 score was selected. This combination was calculated based on the validation set performance.

2.3.4 Explanation Module

The outcome predictions of the machine learning model were explained using the SHAP (SHapley Additive exPlanations) method, which is based on the Shapley value from cooperative game theory *Lundberg and Lee* (2017); *Lundberg et al.* (2018). In game theory, the Shapley value fairly distributes both gains and costs between multiple players with different skill sets in a coalition. Inspired by this concept, in the machine learning context, SHAP values can fairly distribute a predicted probability among input features. This distribution can be either positive or negative. The positive contribution of a variable indicates that it increases the prediction probability, while a negative contribution denotes a reduction in that probability. The Shapley contribution of feature i for patient x is defined as

$$Shapley_i(x) = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [f_{S \cup i}(x_{S \cup i}) - f_S(x_S)] \tag{2.1}$$

where f corresponds to the prediction model to be explained, while F is the complete set of variables, and $S \subseteq F \setminus \{i\}$ corresponds to all possible subsets of F after excluding variable i.

Accordingly, SHAP values enable model interpretation at both the individual patient and population levels. At the patient level, SHAP estimates the contribution of each variable to the predicted outcome. It provides a sense of which variables are contributing to the predicted outcome and to what extent. In the aggregate, the distribution of SHAP values over the whole data set reveals the global behavior of the trained model and input features.

2.3.5 Intelligible Variable Selection Module

2.3.5.1 Step 1 - Statistical analysis to identify variables with robust contributions behaviors

While the global behavior of SHAP values for some variables is robust regardless of the sampled training set, there are variables for which their global contribution distribution vary based on the selected training sample set. For example, the global behavior of creatinine contribution is highly sensitive to the training sample set. As shown in Fig. 2.3 depending on the randomly selected training sample set, the marginal effect of creatinine contribution significantly fluctuates. To identify and exclude variables with non-robust and unreliable SHAP contribution behavior, a bootstrap-based procedure was employed. 1000 bootstrap samples were drawn with replacement from the training set. For each bootstrap sample a separate XGBoost model was trained. Next, the SHAP contribution behavior was estimated for each variable using Kendall's τ correlation coefficient. Kendall's τ is a summary statistic that in this usage assesses the strength and direction of the association between a variable and its SHAP contribution. For each variable, if its correlation coefficient, either positive or negative, is marginally significant with p-value < 0.1, variable is selected to be included in the remainder of the process. The choice of the p-value is arbitrary. In this study, since there is a subsequent variable selection step that examines variable behavior in detail from a clinical perspective, only variables whose behavior was strongly non-significant (p-value > 0.1) were excluded during the statistical inference process.

2.3.5.2 Step 2 - Clinical expert validation of input variables

Once the robust features are selected, the XGBoost classifier is again trained and its predictions are explained using SHAP values. Next, human experts investigated the model explanation in order to complement the machine learning approach. An interdisciplinary team of an expert board-certified physician in neurology and neurocritical care, data scientists, and engineers studied each input variable's contribution to the final outcome prediction. SHAP values of the whole population are used to investigate each variable's marginal effect on the predicted probability. If a variable showed a counterintuitive behavior it was further studied for potential sources of biases. If no logical explanation could be derived, the variable was excluded from the study. Finally, using a robust and medically justified subset of the variables in the XGBoost model, the TBI prognostic model was developed.

2.4 Results

2.4.1 Intelligible Variable Selection Using Computational Analysis and Human Validation

Among 62 candidate variables that were extracted from the EHR (see Table A.1 for the full list of EHR variables and their definitions), 21 were shown to be statistically robust. The robustness was estimated with respect to the global distribution of SHAP contribution. Table 2.2 shows the Kendall's τ correlation coefficients of the robust variables and the corresponding p-values.

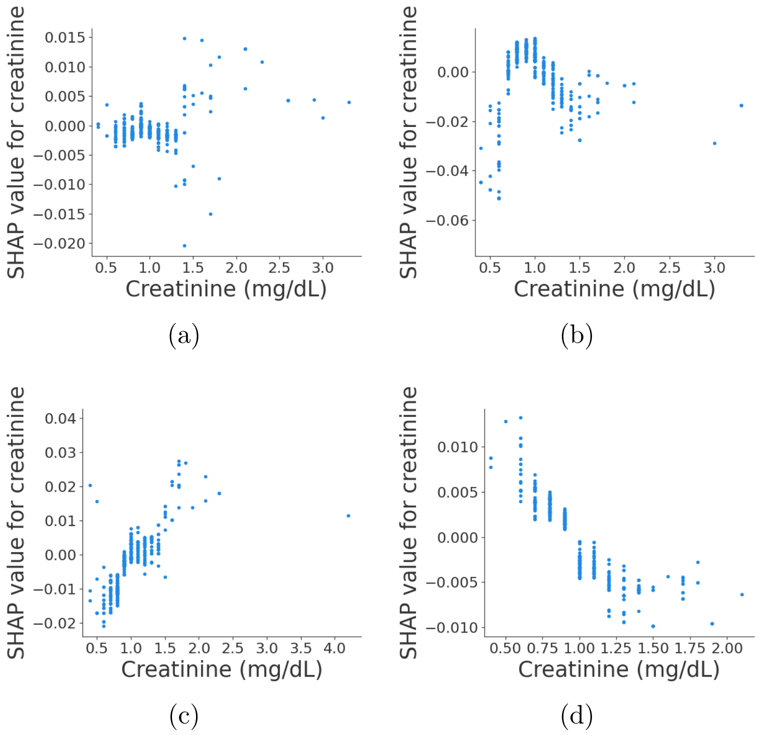
Of the initial 62 features, the 21 selected features are not necessarily those with the highest contributions. For example, creatinine, WBC, and potassium were among the top features in terms of the amount of contribution; however, their behavior was not statistically robust (Fig. 2.3, Fig. A.1, and Fig. A.2).

The 21 automatically selected robust variables where then carefully evaluated to identify those with unexpected or counterintuitive behaviors. Three variables were identified by a physician board-certified in neurology and neurocritical care to exhibit behavior contrary to clinical domain knowledge: active substance abuse (Fig. 2.4a

Table 2.2: The Kendall's τ correlation coefficients and the corresponding p-values of 21 variables that demonstrated robust SHAP global behavior.

Variable Name	**Median**	p**-value**
Age	0.7414	<0.001
Best motor response	-0.7484	<0.001
Best eye opening response	-0.6552	0.003
Best verbal response	-0.7602	0.007
Subdural hematoma (#)	0.6794	0.019
Subdural hematoma (max width)	0.6372	0.018
Subarachnoid hemorrhage (#)	0.7907	<0.001
Intra-ventricular hemorrhage	0.5969	<0.001
Intraparenchymal hematoma (max width)	0.2781	0.066
Brain contusion (#)	0.4964	0.086
Brain contusion (max width)	0.5591	0.018
DAI finding (#)	0.4493	0.025
Third ventricle compression	0.5856	0.012
Transtentorial herniation	0.5200	0.006
Glucose	0.5210	0.054
Hgb	-0.7648	0.003
Platelets	0.6718	0.039
aPTT	0.5862	0.087
INR	0.6223	0.089
Active gastrointestinal disease	-0.3998	0.086
Active substance abuse	-0.6506	0.088

and Fig. 2.5), inactive gastrointestinal disease (Fig. 2.4b), and platelet count (Fig. 2.4c). The results show that the existence of either active substance abuse or active gastrointestinal disease lowers the risk of poor outcomes, which is contrary to clinical domain knowledge. The decrease in platelet count was observed to be associated with better outcomes which also contradicts medical literature (*Maegele*, 2013; *Joseph et al.*, 2014). These counterintuitive associations might be either derived from the collinearity between variables, thereby representing an (un)known proxy variable, or induced by noise in the data set, both of which can introduce bias into the model if not addressed properly. The collinearity effect becomes of crucial importance if it exists between these risk factors and the level of care patients received, which can lead to self-reinforcing positive feedback (*Paulus and Kent*, 2020). For example, patients



(a)

(b)

(c)

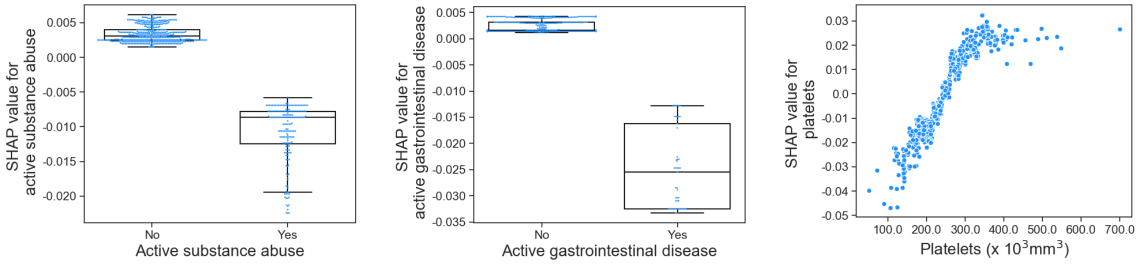
(d)

Figure 2.3: An example of a variable with non-robust contribution behavior. (a), (b), (c), and (d) are the SHAP contribution of creatinine levels to the predicted risk score when training on different bootstrapping sample sets (*Farzaneh et al.*, 2021b).

with active gastrointestinal disease, substance abuse, or coagulation dysfunction (low platelet) might receive more aggressive care that affects their outcome. However, if this bias is not accounted for at the algorithm development level, in real-world settings the model will assess these patients as having a lower chance of unfavorable outcome, resulting in them potentially receiving less aggressive care that they would have otherwise.

These biases are studied in the case study on active substance abuse in the following section. Regarding active gastrointestinal disease and platelet count contributions, no meaningful correlation or explanation was observed in the data set. It is possible that they reflect a latent variable or the observed behavior is merely specific to the study cohort.

These three variables were excluded from the study, leaving 18 EHR variables



(a) (b) (c)

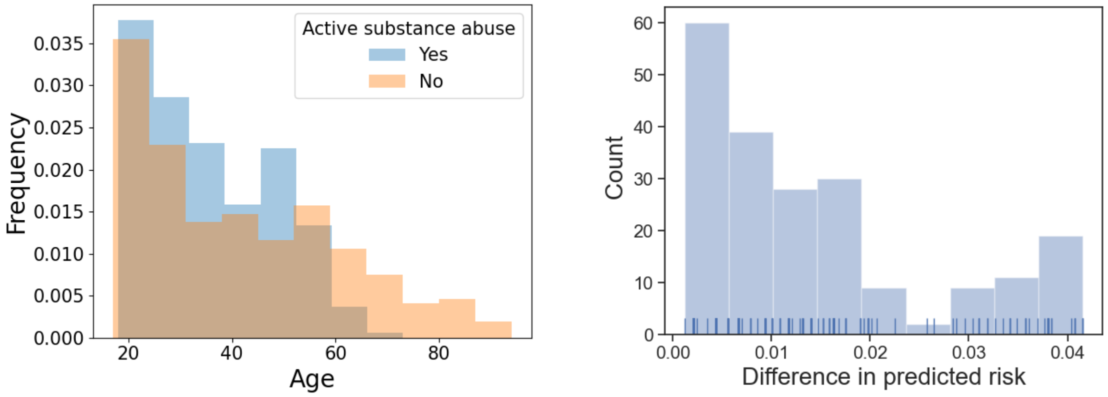
Figure 2.4: SHAP contribution for variable with robust counterintuitive behavior. (a) shows the contribution of active substance abuse. It shows that active substance abuse yields a negative contribution. (b) shows the contribution of presence of active gastrointestinal disease contribution, while (c) shows the contribution of platelet count. In (a) and (b), the center line and box limits correspond to median and upper and lower quartiles, respectively (*Farzaneh et al.*, 2021b).

to be included in the final prognostic model. The selected variables include features from radiology reports, laboratory values, and baseline clinical features (Fig. 2.6a, see Table A.1 for detailed information). Fig. A.3 shows the population level contribution of each of 18 selected variables.

2.4.1.1 Case study: active substance abuse

In this section, the contribution of active substance (alcohol and non-prescribed drug) abuse to risk prediction, along with its possible underlying explanations and potential concerns if not properly addressed, is evaluated. As shown in Fig. 2.4a, active substance abuse negatively contributes to the predicted risk of poor outcome, which is counterintuitive as alcohol and substance abuse are independent predictors of mortality risk.

However, in the study data set, patients with active substance abuse tend to be younger (Fig. 2.5a); thus, this variable might reflect the residual effect of age of a patient. Moreover, the active substance abuse value is correlated with injury etiology, being more common in patients experiencing TBI due to assault ($\rho = 0.19$, p-value < 0.01), and head ISS (Injury Severity Score) ($\rho = -0.12$, p-value < 0.01). Although



(a)

(b)

Figure 2.5: Variables with counterintuitive behaviors. (a) shows the difference in age distribution between active substance abusers and others, (b) shows difference in the predicted risk scores of the test population if the active substance abuse value is set zero or one (*Farzaneh et al.*, 2021b).

no causal correlation can be drawn, we speculate that in the proposed model active substance abuse is a confounding variable due to its simultaneous association with injury severity and assault.

To quantify the effect of active substance abuse on the final prediction, an experiment was performed in which the active substance abuse value was manually set to either zero or one for each test patient. The difference in the predicted risk δp is calculated as

$$\delta p_i = \hat{p}(y_i = 1 | X_{i,i \neq a}, X_a = 0) - \hat{p}(y_i = 1 | X_{i,i \neq a}, X_a = 1), \tag{2.2}$$

where y_i corresponds to the outcome while X_i and X_a are the complete variable set and active substance abuse variable, respectively. Fig. 2.5b shows the histogram of the difference in the predicted risk on the test data set. The average, minimum, and maximum value of δp are 0.014, 0.001, and 0.042, respectively. Based on these results, it can be concluded that in a scenario in which two TBI patients are admitted to a hospital with identical characteristics except for substance abuse, the patient with substance abuse would be predicted to have on average 1.4% (and up to 4.2%) higher chance of a favorable outcome. To address such collinearity-induced biases, variables that exhibited counterintuitive contribution behaviors were excluded from the variable set.

2.4.2 Prognosis of Traumatic Brain Injury Functional Outcome

An XGBoost classifier was implemented to predict the functional outcome - GOSE at 6 months. The classifier was trained, validated, and tested once using the full set of 62 candidate variable, once using 21 identified robust variables, and once using the final 18 intelligible variables (Table 2.3). Although the performance on the training set decreased slightly after excluding non-robust and counterintuitive variables; Area Under the Receiver Operating Curve (AUROC or AUC), accuracy, and F1 score performance on the validation and test sets were well-preserved throughout this process. These results support the conclusion that the excluded features do not affect the performance of the model.

2.4.3 Explaining the Rationale Behind Predicted Risk Scores

The SHAP contribution values provide a detailed view into the risk factors leading to the probability risk score at both the population and individual levels. At the population level, age and the number of brain regions with subarachnoid hemorrhage are by far the most impactful features in determining the elevated risk of poor outcome (Fig. 2.6a). As can be observed in Fig. 2.6a, the contribution of a variable may vary across different patients even if the patients share the same value for that variable. For example, compression of the third ventricle can increase the risk of poor outcome from 3.97% to 6.60% depending on the combination of other risk factors.

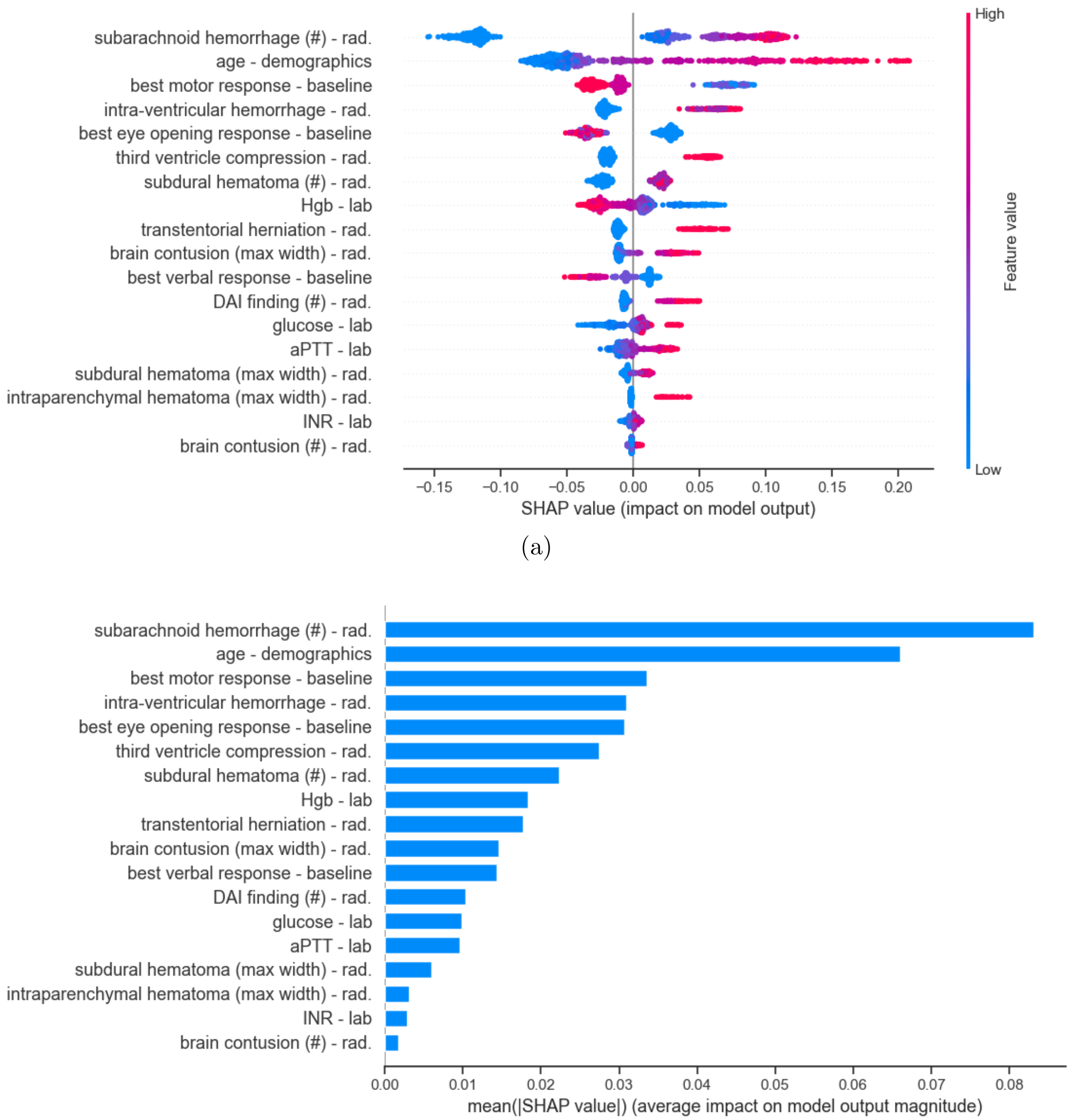
At the individual level, each feature returns a contribution. The aggregate of all feature contributions yields the predicted risk score. For example, for patient shown

Table 2.3: Performance of the TBI prognostic model trained using all candidate variables, only robust variables, and robust and clinically validated variables. Numbers in parentheses are standard deviations. Standard deviation is calculated over 5 cross-validation folds.

All candidate variables			
Sample Set	**Training**	**Validation**	**Test**
AUC (%)	93.72 (2.36)	78.22 (1.26)	80.94
Accuracy (%)	85.22 (3.27)	75.00 (1.69)	75.36
F1 score (%)	82.81 (3.60)	71.29 (1.90)	70.52
Sensitivity (%)	84.77 (3.05)	74.34 (4.89)	70.11
Specificity (%)	85.54 (4.40)	75.49 (4.56)	79.17
Precision (%)	81.06 (5.29)	68.80 (3.30)	70.93
Excluding non-robust variables			
Sample Set	**Training**	**Validation**	**Test**
AUC (%)	90.80 (2.49)	78.77 (1.77)	80.46
Accuracy (%)	81.65 (3.17)	74.84 (2.50)	74.40
F1 score (%)	78.55 (3.44)	71.04 (2.99)	68.64
Sensitivity (%)	80.08 (3.19)	73.94 (5.27)	66.67
Specificity (%)	82.79 (4.50)	75.49 (4.39)	80.00
Precision (%)	77.22 (4.98)	68.62 (3.29)	70.73
Excluding non-robust & counterintuitive variables			
Sample Set	**Training**	**Validation**	**Test**
AUC (%)	89.12 (2.52)	78.36 (1.89)	80.85
Accuracy (%)	80.53 (2.85)	74.51 (2.55)	74.88
F1 score (%)	77.40 (3.75)	70.76 (3.15)	70.45
Sensitivity (%)	80.18 (6.37)	73.93 (5.70)	71.26
Specificity (%)	80.78 (2.38)	74.94 (4.43)	77.50
Precision (%)	75.00 (2.52)	68.13 (3.29)	69.66

in Fig. 2.7a, the presence of subarachnoid hemorrhage in two brain regions increases the predicted risk by 1.84%, while the eye opening response at the time of admission reduces the risk by 4.03%.

The most impactful features for the patient shown in Fig. 2.7a are age, eye opening response, subarachnoid hemorrhage, brain contusion, and subdural hematoma are different from the features that contribute to predicted risk score of the patients shown in Fig. 2.7b.



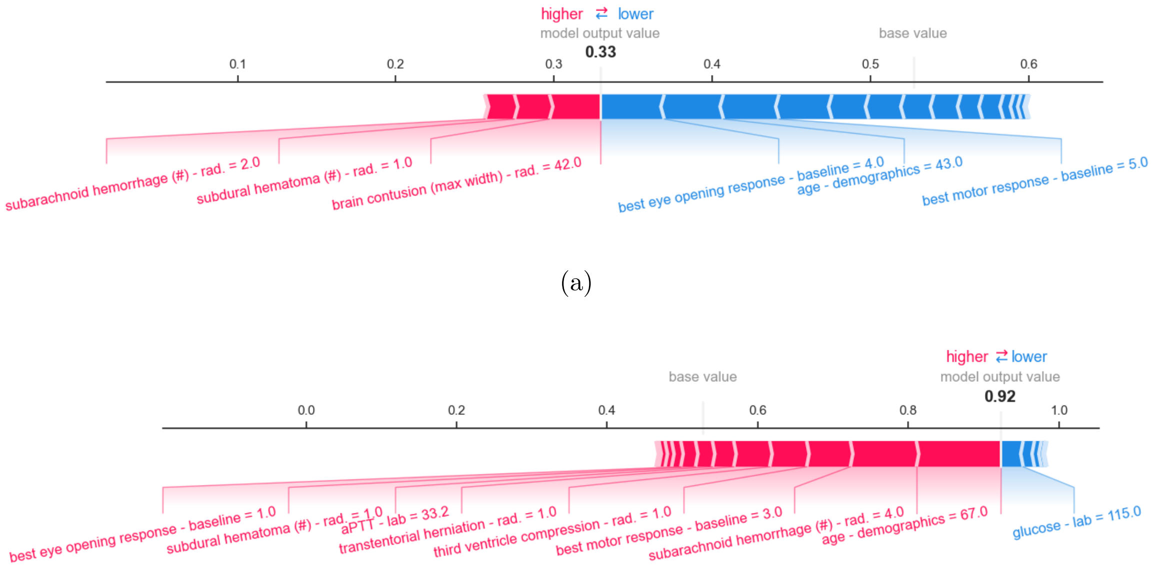
(a)

(b)

Figure 2.6: (a) The summary plot of the contribution of all variables in the final model. Each point corresponds to one patient, while color corresponds to the value of the variable, with the spectrum from blue to pink associated with low to high values. (b) shows variables in order of their importance, where importance of a variable is defined by the average of the absolute SHAP values. Variable types are denoted as *rad*: radiology report and *lab*: laboratory value (*Farzaneh et al.*, 2021b).

2.5 Discussion

This was a secondary analysis of data from the ProTECT III data set, a large clinical trial of patients with moderate and severe TBI. An explainable, expert-guided



(a)

(b)

Figure 2.7: SHAP force plots corresponding to predicted risk scores for individuals (a) and (b). The base value corresponds to the average model output over the training set and is the proportion of the training samples belonging to the class GOSE 1-4. Red and blue arrows, respectively, depict the amount of positive and negative contribution of variables to the predicted risk score. The model output value corresponds to the predicted risk score. For example, the patient shown in plot (a) has a 33% probability of experiencing GOSE 1-4. Variable types are denoted as *rad*: radiology report and *lab*: laboratory value (*Farzaneh et al.*, 2021b).

machine learning framework was developed to automatically predict the long-term functional outcome of TBI patients as defined by GOSE (Fig. 2.1). It is widely acknowledged that transparency and trustworthiness of machine learning models are important factors in real-world applicability, particularly in medical diagnostic and prognostic systems. The proposed framework seeks to move beyond the black box application of machine learning algorithms. Aggregated SHAP values were used to estimate the contribution of each variable to the predicted risk scores at both the population and individual levels. Studying the contributions at the global level enables two rounds of variable selection to be performed, based on: 1) robustness of the contribution of a variable, and 2) clinical domain knowledge.

Among 62 candidate variables from EHR, 21 demonstrated robust global behavior where the global behavior was modeled using Kendall's τ correlation coefficient. Of the 21 robust variables, 3 variables (active substance abuse, active gastrointestinal disease, and platelet count) showed counterintuitive effects on the predicted risk score. Based on the observed behaviors patients with active substance abuse and active gastrointestinal disease were determined to have a better chance of favorable outcome. The lower platelet count was found to be associated with favorable outcomes, which contradicts the clinical literature (*Maegele*, 2013; *Joseph et al.*, 2014). These three variables were excluded from the study as well, leaving 18 robust and clinically validated variables to be included in the final prognostic model.

Finally, an XGBoost classifier was trained to classify patients as having unfavorable (GOSE ≤ 4) or favorable (GOSE > 4) expected outcomes. The final model achieved an AUC, accuracy, and F1 score of 0.8085, 0.7488, and 0.7045, respectively, on the test set. Importantly, the results show that the performance of the model is not negatively affected by reducing the input variable set after imposing the statistical and domain knowledge constraints (Table 2.3). Tree-based models, including XGBoost, are prone to overfitting in the presence of many initial variables. This is evinced in Table 2.3, where the performance of XGBoost on the training set decreased after removing non-robust and counterintuitive variables.

In the final model, age and the total number of brain regions with subarachnoid hemorrhage were the most impactful features in predicting the risk score. These variables were followed by GCS motor score, intra-ventricular hemorrhage, GCS eye opening response, and third ventricle compression. Among laboratory values, hemoglobin, glucose, aPTT (activated Partial Thromboplastin Time), and INR (International Normalized Ratio) were determined to be appropriate predictors.

At the individual level, the model enabled the predicted risk score to be analyzed with respect to which risk factors contributed to a particular decision and to what extent. This tool enables end-users to judge the rationale behind the models' decision making and act accordingly.

To our knowledge, no prior study has used explanatory methods such as SHAP values to intelligibly select variables for black box machine learning classifiers. Using SHAP values for model explanation has become increasingly popular (*Ma and Tourani*, 2020). It is a powerful tool to peer inside black box models and understand how they arrive at a particular decision. Multiple recent studies used only SHAP values for variable selection (*Ma and Tourani*, 2020; *Ogura et al.*, 2020; *Janizek et al.*, 2018; *Bhandari et al.*, 2020; *Bi et al.*, 2020), however, only the variables with the greatest impact as defined by average absolute SHAP value were chosen. This is in contrast to this study, in which it was shown that variables with the greatest contribution are not necessarily robust (Fig. 2.3 and A.2). For example, in the initial model, creatinine level was among the most impactful variables, while through the bootstrapping experiment it was shown to have non-robust behavior (Fig. 2.3). Moreover, the selected high impact features might not align with domain knowledge. For example, in Ogura et al. (*Ogura et al.*, 2020), the total number of traumatic injuries was attributed to a lower risk of death. Thus, our study proposes a novel framework to "intelligibly" select variables using SHAP values, and highlights the importance of collaboration with domain experts.

GOSE at 6-months is the global gold standard functional outcome classification score in TBI prognostication studies. This score is commonly used in major clinical trials, such as ProTECT (*Wright et al.*, 2014) and RescueICP (*Hutchinson et al.*, 2016). However, TBI patient functional outcome can be influenced by non-TBI-related post-injury adverse events. For example, in the test data set, there was a 40-year-old patient that was admitted to the hospital with a 23 mm unilateral subdural hematoma, with no sign of subarachnoid hemorrhage or increased intracranial pressure (e.g., third ventricle compression and transtentorial herniation) indicated in the radiology report. This patient's best motor and eye opening responses were both

4 and within their respective upper quartiles. The patient was predicted to have a 27% chance of poor outcome; however, in reality, the patient died. Looking into post-admission information, the patient developed pneumonia after discharge from the the hospital at day 86 post-injury, leading to the patient's death. This is not the only case of non-TBI-related adverse events experienced post-injury. In the training set there was a 36-year-old subject that, except for 9 mm-wide intraparenchymal hematoma in one brain region, showed no other head abnormalities based on the radiology report. This patient was discharged to home at day 5 post-injury but died of a gun shot at day 87 post-injury. Although it is important, the information about non-TBI adverse events is not recorded for all patients, and even for the patients with such information, it is not mentioned in the data set whether functional GOSE outcome is derived by non-TBI adverse events or TBI alone. To avoid any subjective input, in particular in non-death cases, we did not consider post-injury clinical consolidated comorbidities as an exclusion criteria in this study. Though this limitation of GOSE introduces noise into the ground truth labels that can adversely affect machine learning performance, it is nonetheless the current best proxy for TBI outcomes. It should also be noted that the choice of thresholds, such as a GOSE > 4 being defined as a favorable outcome in this study, is somewhat arbitrary and lacks nuance. Ideally, equally validated but more detailed and objective means to measure TBI outcomes will become available for future studies.

It is also important to acknowledge that there can be "self-fulfilling prophecies" in clinical settings that can influence model performance in ways that are very challenging to mitigate. Self-fulfilling prophecies occur when a perceived poor prognostic factor is present, leading to early withdrawal of care, which then is seen as providing evidence that the prognostic factor is valid (*Williamson and Rajajee*, 2018; *Becker et al.*, 2001; *Hemphill III and White*, 2009).

In addition to imperfect labels, the input variables are susceptible to bias or error.

ProTECT III was conducted at 49 trauma centers in the United States (*Goldstein et al.*, 2017). Given this fact, there exists a potential level of noise or measurement error due to the subjectivity involved in radiology readings, the intrinsic differences in tools for measuring laboratory values, human error during data entry process, among other sources. These measurement errors in EHR can lead to potential loss of predictive power as well (*Duan et al.*, 2016). Given these aforementioned limitations - imperfect labeling, self-fulfilling prophecy, and measurement error - it is important to be cautious when applying models to individual patients.

2.6 Conclusion

In conclusion, we proposed an intelligible machine learning framework for predicting long-term functional outcomes of TBI patients. This framework enables explaining the rationale behind predicted risk scores at both the individual and population level. With the help of a human expert, such an explainable model identifies sources of bias in the prediction model that would not be readily detectable if employed in a black box fashion. Avoiding such biases can ultimately accelerate the adoption of machine learning models in clinical settings.

CHAPTER III

Automated Detection and Severity Assessment of Subdural Hematoma in Traumatic Brain Injury Patients

3.1 Introduction

Subdural hematoma is one of the most common types of traumatic intracranial hemorrhage encountered in neurosurgical practice. It refers to the accumulation of blood in the potential space between the arachnoid mater surrounding the brain and below the skull and dura mater. As the blood accumulates in the subdural region, it can compress the underlying brain parenchyma and lead to focal neurological deficits, unconsciousness and death. Therefore, the existence of SDH may require emergency treatment by surgically evacuating the blood.

The total volume of hematoma is an important factor in diagnosis and prognosis of TBI patients (*Marshall et al.*, 1991, 1992; *Saatman et al.*, 2008; *Gennarelli and Wodzin*, 2008). For instance, widely accepted Marshall scale for TBI severity rating and treatment planning considers the hematoma volume as one of its few key metrics (*Marshall et al.*, 1992). However, manual measurement is time consuming, and almost impossible to implement in practice. Automated image analysis could rapidly measure total hematoma volume and quantify other clinically relevant measurements that are not typically available from a conventional review. Moreover, an automated diagnosis system enables earlier detection, thereby alerting the radiologists for higher prioritization of the imaging study. However, automated detection of SDHs can be challenging due to variability in size, location, and brightness/intensity observed in a head CT scan. The intensity of blood varies substantially over time depending on chronicity of SDH leading to three main types: acute, subacute, and chronic. Acute hematoma is most frequently encountered following severe head injury, and in a head CT scan is identified by its relative brightness (intensity) compared to normal brain tissue. Subacute hematoma develops over weeks, with its brightness similar to normal brain tissue. Chronic hematoma tends to occur in elderly patients with brain atrophy, often as a result of minimal trauma occurring weeks or even months before presentation, and looks darker than normal brain tissue on CT scans.

Automated imaging analysis can act as a safeguard or triage tool for alerting radiologists and clinicians of potentially moderate and severe injuries. Additionally, automated analysis can also quickly measure clinically relevant image features that are difficult to quantify by conventional visual inspection. In the case of SDH, guidelines for surgical intervention and many treatment decisions are based on simple measurements of maximum hematoma width at a single CT slice. Hematoma volume is not typically measured or reported because the accurate measurement is difficult and time-consuming. Automated image analysis could rapidly measure total hematoma volume and quantify other clinically relevant measurements that are not typically available from a conventional review. Since subdural hematomas frequently expand and sometimes require multiple surgical interventions, another application can be using quantitative volumetric measurements to accurately assess both progression of disease and response to surgical intervention.

Consequently, the goal of this study is to design an automated platform to detect SDH and assess its severity by measuring its volume in patients with potential TBI.

Moreover, we perform an inter-physician variability analysis to determine a human performance benchmark. The resulting algorithm could enable timely intervention, thereby improving survival while lowering the risk of secondary injuries.

For the SDH segmentation task, classical image processing approaches were integrated with a deep convolutional neural network model to overcome the limitations of each method. The proposed algorithm employs a joint feature representation of domain knowledge-driven hand-crafted features and data-driven deep features to train a random forest model.

3.2 Related Works

To the best of our knowledge, there is no prior published work that describes automated methods to detect and segment the different types of SDH. However, there are several techniques developed for acute hematoma detection/segmentation which are reviewed here. These techniques fall into semi- or fully-automated categories regarding human interaction. These segmentation techniques either implement traditional rule-based methods (*Yuh et al.*, 2008; *Liao et al.*, 2009, 2010; *Chan*, 2007; *Liu et al.*, 2008) or employ machine learning methods (*Bhadauria et al.*, 2013; *Shahangian and Pourghassem*, 2013, 2016; *Grewal et al.*, 2018; *Chilamkurthy et al.*, 2018).

Yuh et al. (*Yuh et al.*, 2008) evaluate the presence or absence of acute intracranial blood from CT scans in an automated fashion. Their findings include the presence or absence of: (1) SDH or epidural hematoma, (2) subarachnoid hemorrhage, and (3) intraparenchymal hematoma. First, regions with intensity similar to blood are detected using thresholding. Then, the detected potential region is categorized into one of the above three categories based on its location, size, and shape. If a blood cluster is contiguous to the skull, it is defined to be SDH or epidural hematoma. Otherwise, it is further divided into subarachnoid or intraparenchymal based on shape and location. However, this study only accounts for acute hematoma. Liao et al.

(*Liao et al.*, 2009) focus on hematoma in different brain locations. In their method, all patient images with subacute (with the same brightness as normal brain tissue), and/or chronic (darker than normal brain tissue) hematoma regions are manually excluded. For other images, they focus on a single pre-selected CT slice containing the largest intracranial area. Within that slice, the largest hyperdense component is found. Next, a level set method is applied to evolve the segmentation. In a more recent study (*Liao et al.*, 2010), the same research group proposed a multiresolution binary level set method to identify the hematoma for patients with neurological disorders. As in their previous work, only acute hematoma is considered. Moreover, all of the included CT images are from patients who underwent brain surgery. These restrictions result in selected slices having a large portion of blood, making detection less challenging. Chan et al. (*Chan*, 2007) introduce a method to segment small acute brain hematoma. This method first detects bright objects by simple thresholding and then searches for right-left asymmetry of brain tissue to select candidates. Finally, acute hematoma regions are identified based on both anatomical location and image features. The proposed symmetry analysis, however, is highly dependent on the fact that axial slices are perfectly parallel to the axial plane, which is not always the case. The method proposed in Li et al. (*Liu et al.*, 2008) aims to segment acute intracranial hemorrhage. In this method, an information gain algorithm is implemented to adaptively determine the hematoma intensity interval. In this model, it is assumed that all the images have hematoma, negating the need to identify non-hematoma images.

A Fuzzy C-Means (FCM) clustering and active contour model are employed in Bhadauria et al. (*Bhadauria et al.*, 2013) to segment acute hematoma. First, FCM is used to group the brain tissue based on image intensity. An active contour model is then applied to refine the contours of the clusters detected by FCM. The number of clusters in the FCM algorithm must be predefined; however, this number is subject to change based on the presence/absence of hematoma and also other types of TBIs.

In Shahangian et al.(*Shahangian and Pourghassem*, 2013) a fixed threshold is applied on images to approximate acute hemorrhage regions. Then, multiple textural and geometrical features are extracted from a selected region to classify it as intracranial, epidural hematoma, or SDH. This work is further improved in Shahangian et al. (*Shahangian and Pourghassem*, 2016) by applying an adaptive threshold instead of a fixed value. There are a few research studies related to brain hematoma detection that employ convolutional neural networks. The focus of (*Grewal et al.*, 2018; *Chilamkurthy et al.*, 2018) is to determine the presence or absence of acute hematoma in CT images without investigating the hematoma segmentation and severity assessment.

One problem with most of the aforementioned studies is that they narrowly focus on anatomic location and size of hematoma. For instance, only hematoma thickness larger than 4 mm is considered in Liao et al. (*Liao et al.*, 2009). Additionally, none of the referenced studies focus on different types of hematoma, even though subacute and chronic SDHs are widely presented in specific patient populations such as in the elderly, who frequently present with a combination of acute and chronic hematomas-Moreover, for the rule-based models, algorithmic hyperparameters are often manually selected by analyzing the whole data set, rather than just the training set, potentially producing a model overfitted to the particular data set. Additionally, except for Bardera et al. (*Bardera et al.*, 2009), no other work takes advantage of available 3D information. The use of 2D images alone fails to consider spatial coherency that can help avoid false positive or false negative regions.

In our preliminary work (*Farzaneh et al.*, 2017b), we reported on 35 CT scans included in the current study. The current study expands on this by including a much larger sample size, substantially improving the segmentation algorithm by integrating domain knowledge into data-driven deep models, performing further analyses, and comparing performance with a human benchmark.

3.3 Materials

3.3.1 Patient Population

This HIPAA Compliant, Institutional Review Board (IRB) approved retrospective study utilized a data set of TBI patients aged 18 years or older admitted to the University of Michigan Neurological Intensive Care Unit (ICU) or Emergency Department from January 1, 2010, to August 10, 2015. In this study, the written informed consent from patients is waived by IRB because the research involves no more than minimal risk to the subjects. The only risk to subjects is breach of confidentiality and loss of privacy. Safeguards were in place to prevent this from happening. Moreover, research could not practically be carried out without the waiver or alteration because this study exclusively involved the secondary analysis of information that was collected as a part of routine medical care. A number of patients to be included in the study could have been died or lost to follow-up so could not conceivably be contacted to provide consent. We included 98 consecutive patients with SDH at the cerebral convexities as well as 12 subjects with normal head CT scans. We did not include interfalcine hematomas in the midline.

3.3.2 CT Image Setting

All the images used to develop our method were acquired in the axial plane on 64 slice CT scanners (either LightSpeed VCT or Discovery CT HD750, GE Healthcare Milwaukee, USA). Images were stored on a McKesson Picture Archiving and Communication System (PACS). All CT scans were acquired with 0.625 mm thickness and reformatted to 5 mm slice thickness for storage purpose while the pixel spacing in the axial plane ranges from 0.4297 mm to 0.4883 mm.

3.3.3 CT Image Annotation

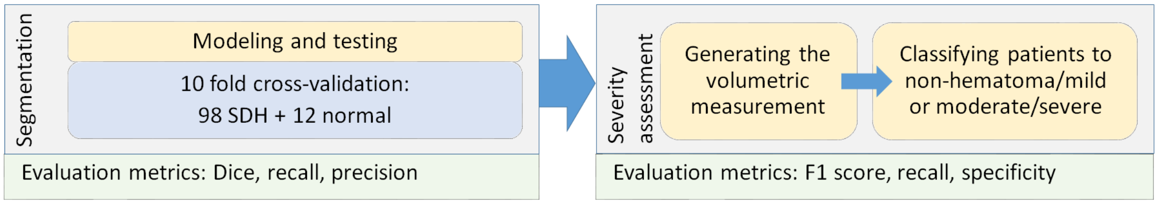
In order to train our model, we used images annotated by two skilled fellowship trained neuroradiologists with 29 and 16 years of experience respectively. The "Free Hand" tool of the MicroDicom DICOM viewer software was employed to annotate CT scans (*MicroDicom*, Accessed: May 1, 2020). Both neuroradiologists were blinded to other information. Although annotating images from multiple experts could minimize inter-physician variability and thereby improve label accuracy, performing this process is labor-intensive and prevents creation of large training set. Therefore, in our study the data set was divided into two parts and each part was annotated by one of the neuroradiologists. All annotations were then reviewed and adjudicated by an expert board-certified physician in neurology and neurocritical care. If any annotation done by one of the neuroradiologists appeared to include a potential human error (as adjudicated by the neuro ICU physician), the other neuroradiologist was asked to review and revise the annotation if indicated. In this way, annotation uniformity was optimized and improved. The adjudicated annotations are used as the ground truth throughout the chapter. A subset of 20 CT scans was annotated by both neuroradiologists to determine the inter-physician variability.

3.4 Methods

3.4.1 Subdural Hematoma Assessment Framework and Experimental Design

The study design for SDH segmentation and severity assessment are shown in Fig. 3.1. We first developed a machine learning model using 10 fold cross-validation to segment SDH regions. Then, we employed the automatically segmented region to assess the severity of SDH.

Fig. 3.2 demonstrates a high-level overview of the proposed method. Given a



Segmentation
Modeling and testing
10 fold cross-validation:
98 SDH + 12 normal
Evaluation metrics: Dice, recall, precision
Severity assessment
Generating the volumetric measurement
Classifying patients to non-hematoma/mild or moderate/severe
Evaluation metrics: F1 score, recall, specificity

Figure 3.1: Experimental design and evaluation strategy for SDH segmentation and severity assessment (*Farzaneh et al.*, 2020).

head CT scan, we first performed pre-processing to identify the region of interest (ROI). Next, potential discriminative patterns were extracted from the ROI, followed by using a random forest classifier to identify SDH regions. Finally, in the post-processing step, morphological operations and Gaussian kernel smoothing were used to improve overall segmentation performance.

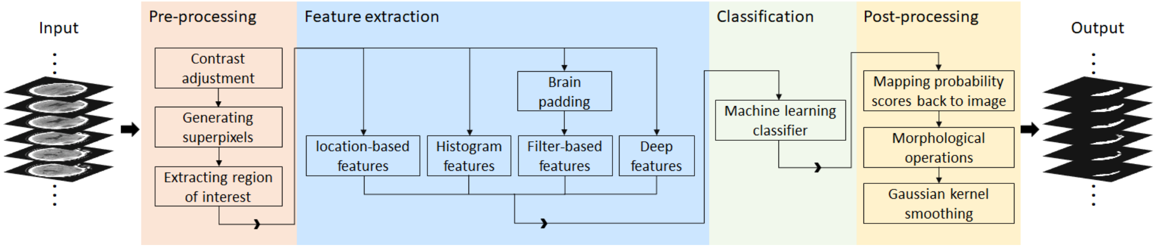


Figure 3.2: A schematic diagram of the proposed SDH segmentation method (*Farzaneh et al.*, 2020).

3.4.2 Subdural Hematoma Segmentation

3.4.2.1 Pre-processing

Pre-processing began with the generation of a 3D representation of the CT slices from a sequence of 2D images, and was followed by intensity normalization of images according to the provided metadata, skull segmentation, and extracting ROI. Then, intracranial pixels (i.e., pixels enclosed by the skull) were grouped into superpixels that were used throughout the feature extraction and classification steps.

Contrast Adjustment and Skull Segmentation

First, the brain tissue intensity was normalized with respect to the CT metadata parameters including window center and window width. Although these parameters are used to adjust and normalize the intensity of normal brain tissue, they may not cover the pathologic tissue intensity range. To solve this problem, the window width, ww, was expanded to cover a larger range of pixel values as:

$$ww_{new} = \alpha \times ww_{old} \tag{3.1}$$

where $\alpha > 1$.

To normalize the pixel intensities, we performed a linear mapping shown in (4.3) to adjust the contrast based on ww_{new}, and window center, wc.

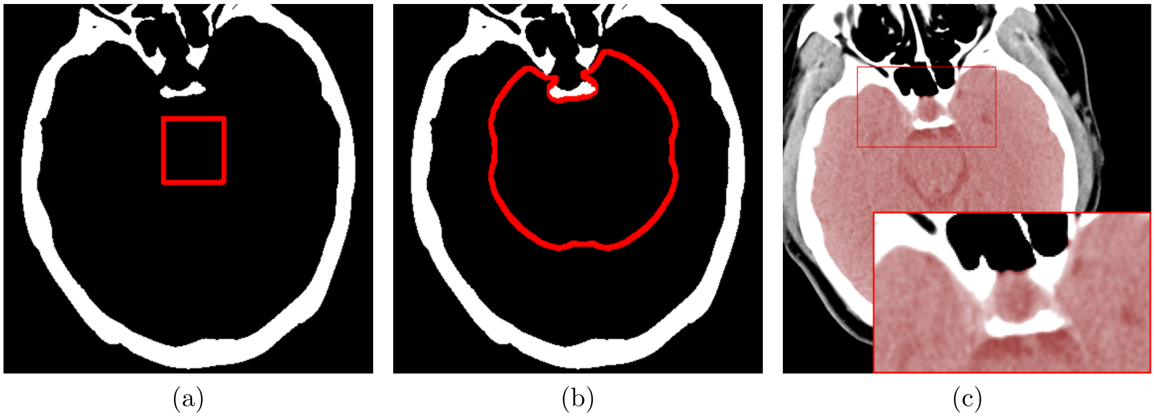
$$I_{ca}(i,j) = \begin{cases} 0 & I(i,j) < a \\ I_{max} & I(i,j) > b \\ \frac{I(i,j)-a}{ww_{new}} \times I_{max} & O.W. \end{cases} \tag{3.2}$$

In (4.3), $a = wc - \frac{ww_{new}}{2}$ and $b = wc + \frac{ww_{new}}{2}$. I and I_{ca} correspond to input and contrast adjusted images, respectively. According to (4.3), the range of $[a,b]$ was mapped to the range of $[0, I_{max}]$ where I_{max} was set to 255. Since ww_{new} was lower than I_{max} no textural information was lost by this method of mapping.

Next, the skull was segmented with respect to its Hounsfield unit. The range of pixel values belonging to each part of the body can be identified by its radiodensity which is quantified according to Hounsfield scaling parameters.

Extracting the region of interest

In our database, even the most severe case of SDH was not deeper than 3.2 cm from the skull. Therefore, the ROI was defined as the intracranial region within 3.2 cm of the inner skull. A level-set method was employed to segment the intracranial region enclosed by the skull in case there were any openings in the skull boundary (Fig. 3.3). These openings can exist due to either normal anatomy such as the eye cavity or traumatic injury such as a fracture. If the skull was closed in the axial plane, simply, all pixels enclosed by it were selected as the intracranial region.


(a)

(b)

(c)

Figure 3.3: Intracranial region segmentation by implementing the level-set method in skulls that have openings. (a) Initial region, (b) modified output after 20 iterations, (c) final result after 50 iterations; the transparent red region indicates our segmented inner skull region (*Farzaneh et al.*, 2017b)

.

Sampling

Once the ROI was delineated, the image was divided into non-overlapping regions of connected pixels with approximately similar gray value. Superpixels were used instead of pixels to reduce redundant information. In this work, we used the Simple Linear Iterative Clustering (SLIC) algorithm (*Achanta et al.*, 2012) for generating superpixels.

3.4.2.2 Feature Extraction

Feature extraction was performed to derive patterns and statistics that represent superpixels. These features were later used as inputs to the machine learning classifiers to discriminate SDH superpixels from non-SDH ones. In this study, we considered hand-crafted textural and spatial features, as well as deep image features. Additionally, the patient's age is incorporated as an auxiliary variable into the machine learning algorithms. Age-related brain atrophy predisposes patients to chronic SDH formation and can also lead to subdural hygroma, the accumulation of cerebrospinal fluid within the skull, which can be mistaken for chronic SDH. A full list of extracted features can be found in Table B.1.

Histogram and filter analyses were performed to extract hand-crafted local textural information from the images. In order to derive local appearance information, a window with a fixed size of 25×25 pixels was localized around the center of each superpixel from which the corresponding features were derived. However, if the window was centered on superpixels close to the skull, where SDH tends to occur, the selected window would include pixels from skull. Thus, extracted textural features from the window would be affected by skull pixels and might not correctly represent SDH characteristics. For example, features that normally represent the orientation of brain/hematoma texture, would instead represent the skull orientation, which is not informative. Since SDH tends to occur adjacent to the skull extracting correct features in this region is especially important for our purposes. In order to deal with this challenge, we removed the skull from the image and padded the image with a symmetric mirror reflection across the inner surface of the skull. As the boundary of the brain is not a straight line, we could not employ standard padding techniques. Instead, we proposed to iteratively pad the image across the irregular border of the intracranial region. This task was performed by maintaining two masks: the inner mask that shrank in each iteration, and the outer mask that grew in each iteration.

The region grown during each iteration was filled using the nearest pixel on the inner mask. Fig. 3.4 illustrates the intracranial region before and after padding, as well as the corresponding filtered image by applying a Gabor filter. The used Gabor filter highlights the textural component in the $\frac{3\pi}{4}$ orientation. In Fig. 3.4b, though there is no significant textural component in the $\frac{3\pi}{4}$ orientation, the filter enhances the border incorrectly within the region enclosed in yellow. As shown in Fig. 3.4d, this issue was resolved using the proposed padding method.

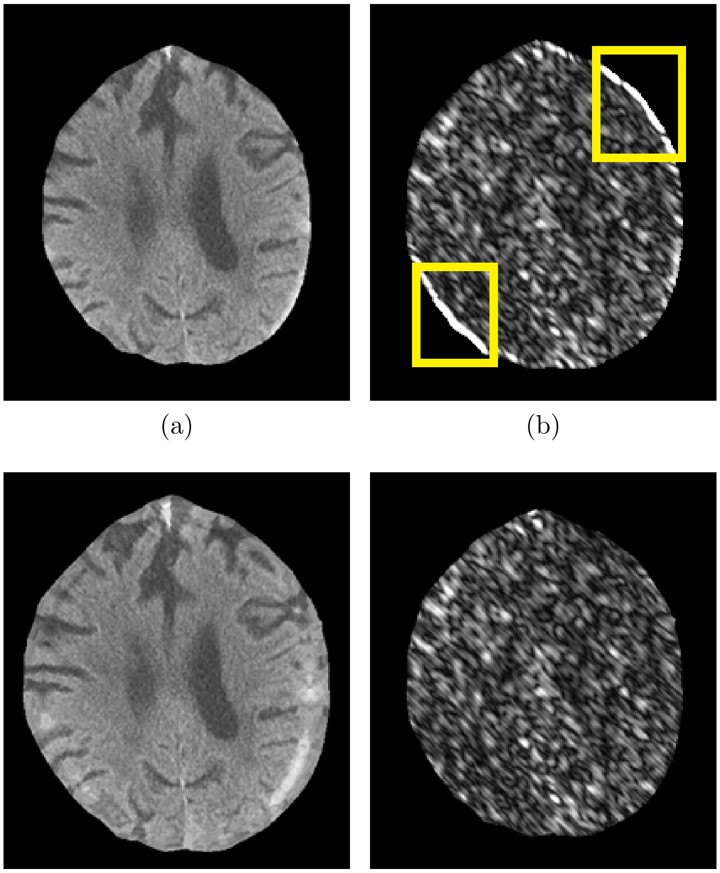


Figure 3.4: The effect of padding on a Gabor descriptor. (a) The original image before padding, (b) the magnitude response for a Gabor filter at $\frac{3\pi}{4}$, (c) the image after applying the proposed padding method, (d) the magnitude response for the Gabor filter at $\frac{3\pi}{4}$ applied on (c) (*Farzaneh et al.*, 2020).

Histogram-based Statistical Features

For each superpixel, a histogram of pixel intensity inside the corresponding window was generated. The histogram-derived features included minimum, maximum, average, and standard deviation, σ, of pixel intensity. The average intensity of pixels within the superpixel was also calculated. Other features were skewness, kurtosis, entropy, and smoothness. There are various definitions of "smoothness". In our approach, smoothness was calculated using Equation (3.3).

$$smoothness = 1 - \frac{1}{1-\sigma^2}. \tag{3.3}$$

Filtering-based Features

To integrate more textural information, a group of features was extracted by convolving Gabor and Laplacian of Gaussian filters with the images.

Gabor filters were used to extract image features at multiple orientations and frequencies. The Gabor filter was calculated by Equation (3.4), where u_0 is the frequency of a sinusoidal carrier along the x-axis, and σ_x and σ_y are respectively the constant of the Gaussian envelope along the x and y axes. To build the filter for orientations other than 0° a rigid rotation of the x-y coordinate was performed (*Jain and Farrokhnia*, 1991).

$$h(x,y) = \exp\left(-\frac{1}{2}\left[\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2}\right]\right) \cdot \cos(2\pi u_0 x) \tag{3.4}$$

In this study, we used Gabor filters in 8 evenly spaced orientations $\{0, \frac{\pi}{8}, ..., \frac{7\pi}{8}\}$ (Fig. 3.5), each at 4 wavelengths of the sinusoidal carrier ranging from 2 to 16 pixels/cycle. Each of these filters detects the edges along the particular direction and we used the median of a filtered window as a feature in our model.

The Laplacian of Gaussian filter determines the edge content of the images in a lo-

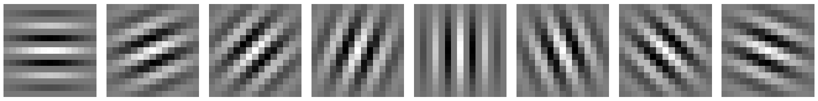


Figure 3.5: Real parts of Gabor filter in 8 implemented orientations (*Farzaneh et al.*, 2017b).

calized region. This filter measures the second spatial derivative of a smoothed image and highlights the edges. Gaussian filters with five different kernels were implemented to smooth the image.

Location-based Features

As mentioned above, subdural hematomas are more likely to occur in specific regions. Thus, location information was incorporated into our model by extracting features in a spherical coordinate system. These parameters are radial distance, azimuth angle, and elevation angle. The origin was fixed on the center of the mass of the skull on the lowest selected slice. As shown in Fig. 3.6, the elevation angle, φ, determines the angle between the horizontal plane and the line connecting the point to the origin. The azimuth angle, θ, corresponds to the angle between the projection of the line connecting the point to the origin on the horizontal plane and the brain midline that separates it into two hemispheres. The radial distance, r, measures the distance of the point to the origin. As proximity of a point to the skull increases the likelihood of injury, in addition to the radial distance, we considered the distance between the point and the inner surface of the skull, r'.

Deep Features

In this work, the U-net architecture was employed to extract data-driven deep features. The U-net architecture was proposed by Ronneberger et al. (*Ronneberger et al.*, 2015) in 2015 and has subsequently become one of the most widely used convolutional neural network architectures for biomedical applications, specifically those

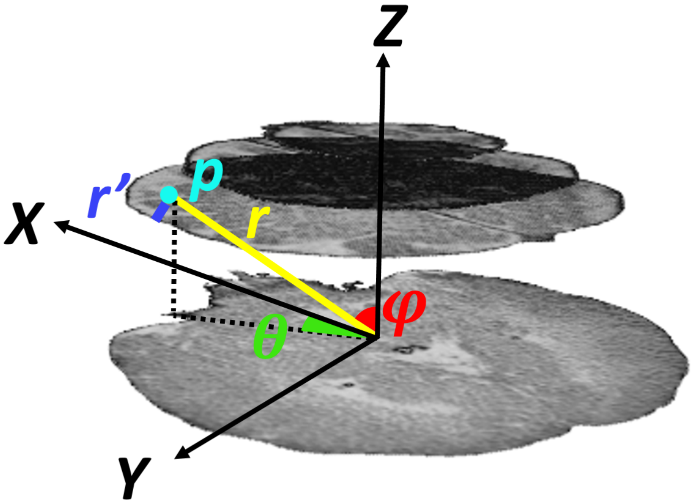


Figure 3.6: Illustration of location-based features of point p within the region of interest (*Farzaneh et al.*, 2020).

with limited access to annotated data. First, a baseline U-net model shown in Fig. 3.7 was trained for SDH segmentation and then activations of the second layer before the output layer were employed as deep features. The layer used for deep feature extraction is marked by blue in Fig. 3.7 and yields 64 features for each superpixel. To ensure that deep features were not influenced by the test set, n-fold cross-validation was performed, in which n different models generated n sets deep features. Later, the same cross-validation folds were used to train the random forest model.

3.4.2.3 Classification

Once the aforementioned features were extracted for each superpixel, a random forest model was trained to classify each superpixel sample as hematoma or non-hematoma. As the number of negative sample points (superpixels) was approximately eight times more than positive sample points, the classes were highly imbalanced. To overcome this challenge, undersampling on the negative samples was performed, in which all positive points were kept and an equal number of points from the negative class were randomly selected to be used in the training phase.

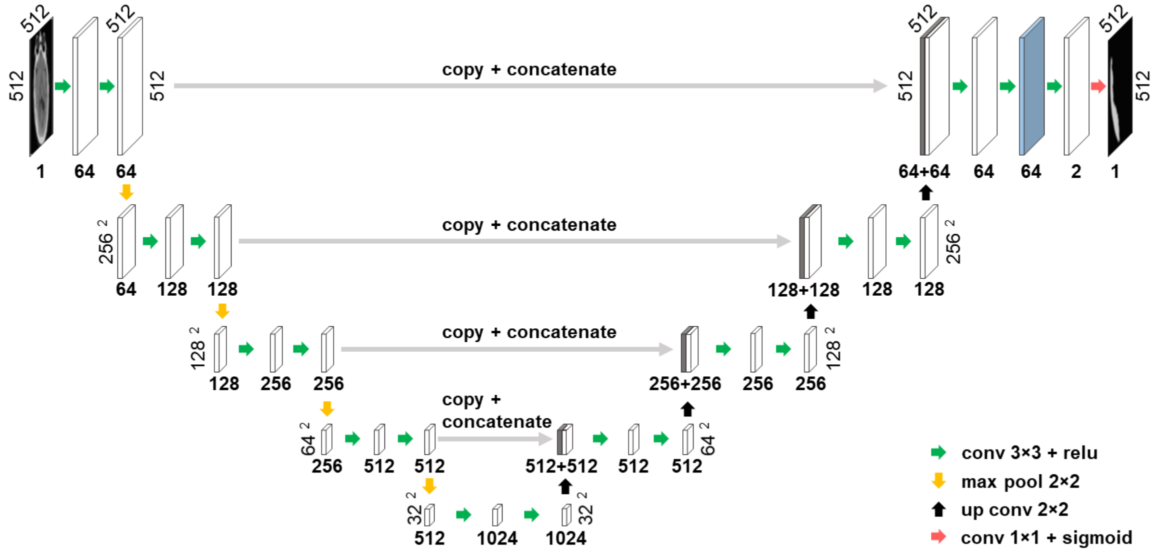


Figure 3.7: Illustration of the baseline U-net architecture used for segmentation. The activations of the layer marked in blue are used as deep features (*Farzaneh et al.*, 2020).

3.4.2.4 Post-processing

Using the label calculated for each superpixel in the classification step, the initial SDH mask was formed. However, this initial SDH mask was generated without accounting for information in neighboring pixels.

Multiple disjoint subdural hematomas may occur simultaneously. However, pixels within each distinct hematoma must be connected. This contextual information was ignored during the previous superpixel classification step. As such, there might be misclassified superpixels that can be corrected by considering their neighboring regions. In order to resolve this issue, small sparse components were first excluded from the initial 2D SDH mask. Next, morphological operations were applied to fill the small holes and gaps between the segmented mask and the skull. Additionally, 2D masks disconnected within the third dimension were excluded. In addition to applying morphological operations, the segmented mask was filtered using a 3D Gaussian smoothing kernel to smooth jagged contours and increase spatial coherency.

3.4.3 Hematoma Volumetric Measurement for Severity Assessment

Hematoma volume is a key metric used in TBI severity ratings, such as the widely used Marshall scale(*Marshall et al.*, 1992). However, because its manual measurement is not easy, subdural hematoma width rather than total volume is much more routinely used in clinical practice. For the purpose of severity assessment, we categorized hematoma severity based upon volume: < 25 cc - non-hematoma/mild; and ≥ 25 cc - moderate/severe. To quantify the hematoma volume, we used the segmented SDH regions as a binary mask, and calculated the volume of lesion using the CT slice resolution from the CT scan metadata.

3.4.4 Inter-physician Variability Analysis

To create a benchmark with which to evaluate segmentation performance, we sought to determine the inter-physician variability of subdural hematoma annotation. In this analysis, a set of 20 SDH patients representing the same volumetric distribution as the original 98 patients were selected. We asked two skilled neuroradiologists to independently annotate the selected set. The Dice similarity coefficient was used to compare the two sets of annotations.

3.5 Results

3.5.1 Classification Result

To train the classifiers, the data set of 110 patients was partitioned into 10 folds for cross-validation. The folds were generated to include a roughly balanced distribution of SDH types. The random forest yielded the average AUC of 0.9777 in classifying superpixels using the combination of hand-crafted and deep features.

3.5.2 Subdural Hematoma Segmentation

Table 3.1 compares the segmented SDH region with the ground truth. Dice is a summary measure of spatial overlap between the segmented region and the ground truth. Dice, sensitivity, precision, and specificity evaluation metrics are defined as:

$$\begin{aligned}\text{Dice} &= \frac{2 \times |S \cap GT|}{|S| + |GT|}\\ \text{Recall} &= \frac{|S \cup GT|}{|GT|}\\ \text{Precision} &= \frac{|S \cup GT|}{|S|}\\ \text{Specificity} &= \frac{|\overline{S} \cap \overline{GT}|}{|\overline{GT}|}\end{aligned} \tag{3.5}$$

where GT and S refer to the manually annotated ground truth and the algorithm segmentation, respectively. $|\,.\,|$ denotes set cardinality and $\overline{A}$ is the complement of A.

Table 3.1 shows that applying post-processing increased the average Dice and decreased its variation. Since there was no SDH and hence no ground truth region annotated for patients without SDH, Dice, recall, and precision values were not defined for such subjects and therefore they are not included in the table. For these subjects, specificity was calculated to determine what portion of negative pixels within the defined ROI was correctly classified. For healthy subjects, our algorithm reached an average specificity (± standard deviation) value of 99.89 (±0.36)%.

Table 3.1: A comparison of the mean of the segmentation performance metrics before and after post-processing as well as the performance of the baseline U-net architecture. Numbers in parentheses are standard deviations.

Performance Metric	**Proposed Method without Post-processing**	**Proposed Method**	**U-net**
Dice (%)	73.63 (16.20)	75.35 (15.62)	73.20 (19.98)
Recall (%)	80.72 (16.75)	78.61 (19.09)	70.93 (21.31)
Precision (%)	70.75 (17.40)	76.12 (15.04)	79.61 (20.10)

It is noteworthy that in Table 3.1, the standard deviations of both the proposed model and U-net are calculated at the pixel-level and not at the patient-level. The results of the proposed segmentation algorithm at different stages are shown in Fig. 3.8 and 3.9. These images cover various combinations of size and type of subdural hematoma.

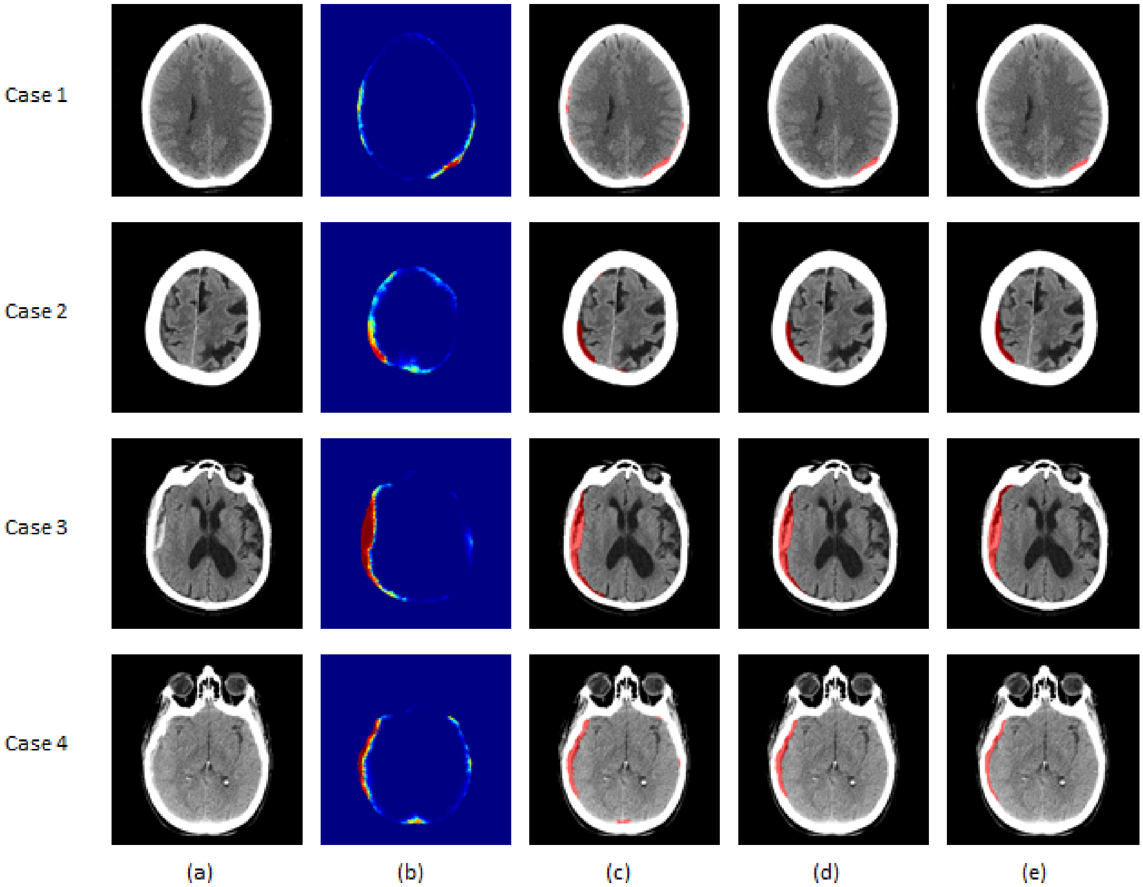


Figure 3.8: SDH segmentation results on mild and moderate subjects. (a) The original CT image, (b) the probability map corresponding to the output of the classifier, (c) the segmented region before post-processing, (d) the segmentation result after post-processing, (e) the ground truth. Cases 1 and 2 are mild (< 25 cc of blood) patients, and cases 3 and 4 are moderate (25-50 cc of blood) ones. Cases 1, 3 and 4 are acute SDH while case 2 is an example of chronic hematoma (*Farzaneh et al.*, 2020).

Next, the generalizability of the proposed segmentation approach was further validated by analyzing segmentation performance with respect to different ranges of hematoma volume (Fig. 3.10). The reference volume was calculated using the ground truth, and discretized by thresholding at 25, 50, 100, and 200 cc. It can be concluded

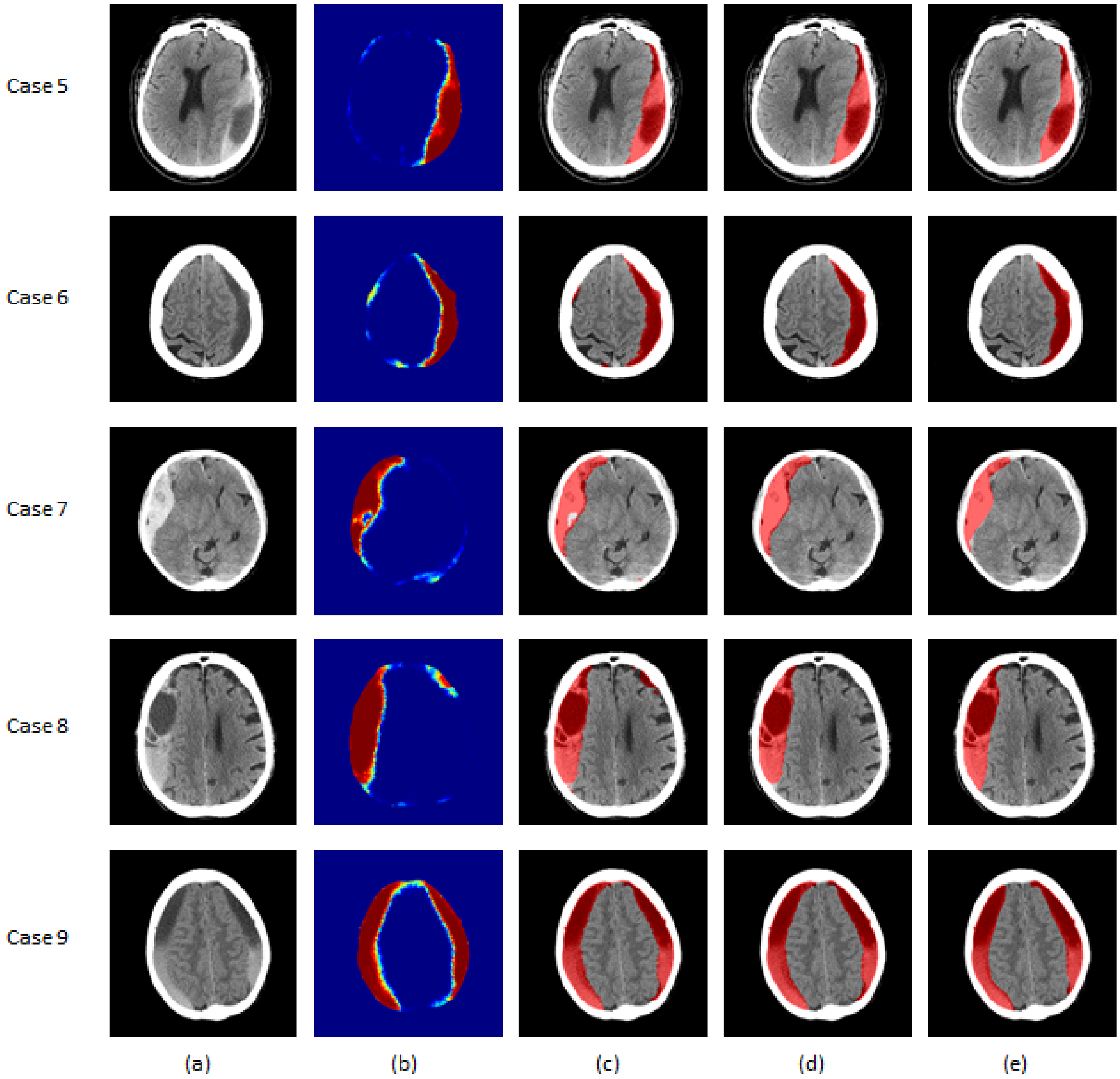


Figure 3.9: SDH segmentation results on severe (> 50 cc of blood) subjects. (a) The original CT image, (b) the probability map corresponding to the output of the classifier, (c) the segmented region before post-processing, (d) the segmentation result after post-processing, (e) the ground truth. Cases 5 and 6 are severe patients with total blood volume of less than 100 cc, while cases 7 and 8 have total blood volume of 100 to 200 cc, finally case 9's SDH volume is over 200 cc. Cases 5, 8 and 9 contain a mix of acute and chronic hematoma. Case 6 is an example of chronic hematoma while case 7 is acute (*Farzaneh et al.*, 2020).

from Fig. 3.10 that Dice values for subjects with mild SDH (47.67%) were lower than for those with more severe SDH (79.97%), possibly because the smaller regions were less represented in the classifier. Moreover, for small lesions, even a small deviation in segmentation has a large impact on the Dice similarity value.

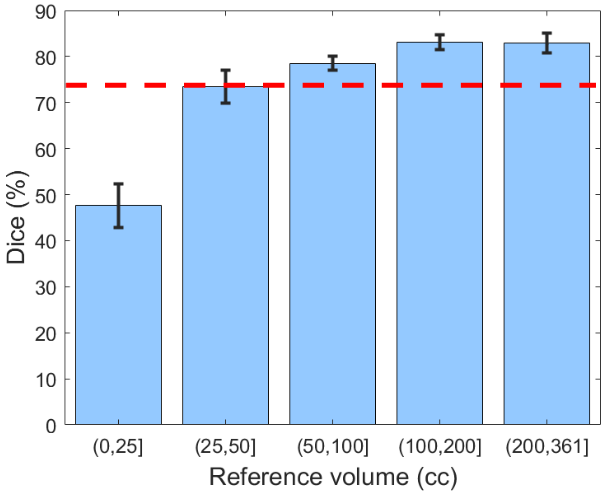


Figure 3.10: The Dice similarity coefficient based on the severity of hematoma. Error bars represent ± 1 Standard Error of the Mean (SE), the 68% confidence interval. Red dashed line indicates the human benchmark (*Farzaneh et al.*, 2020).

Table 3.1 also provides a comparison between the performance of the proposed method and the U-net model. To ensure a fair comparison, both models were trained using 10 fold cross-validation. As shown in Table 3.1 the proposed method outperforms U-net in terms of both the average Dice value and the standard deviation. Moreover, the summary statistics in Fig. 3.11 evince that compared to U-net, the proposed method yields less variability for all severity levels. Likewise, the lower extreme of the proposed method is greater than the lower extreme of U-net in all categories. In particular, for mild hematomas with less than 25 cc of blood, more than 25% of U-net's Dice values are lower than any of those from the proposed method. The lower average performance and higher variability of U-net may be because deep learning approaches require a large and representative sample of annotated images. Thus, a deep learning model trained on limited data sets may fail to reflect the unseen spectrum of the real-world data distribution. Integrating hand-crafted features that reflect human domain knowledge can compensate for this limitation of deep models, yielding segmentation results that are more consistent and robust.

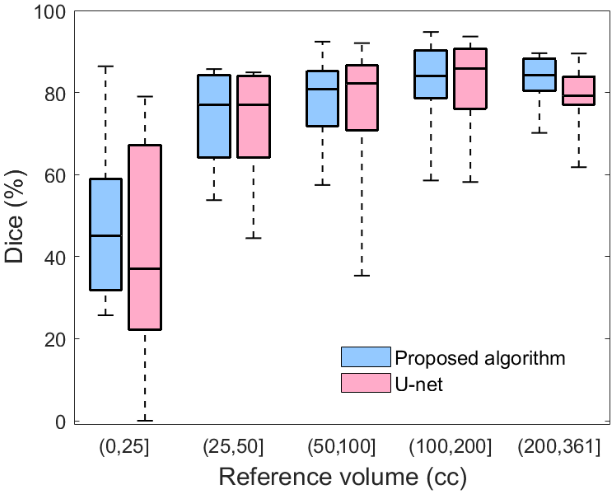


Figure 3.11: A box plot comparing the Dice similarity coefficients of the proposed algorithm and U-net with respect to the severity of hematoma (*Farzaneh et al.*, 2020).

3.5.3 Generalizability of the Segmentation Model over Different Types of SDH

We investigated the performance of the proposed algorithm with respect to subdural hematoma type (Table 3.2). We used the metrics defined in Equation (4.3) to compare the algorithm's segmentation results with the manually annotated ground truths.

Table 3.2: Mean of the segmentation performance metrics for acute, subacute and chronic types, and their mixture. Numbers in parentheses are standard deviations.

Performance Metric	**Acute (n=37)**	**Subacute (n=4)**	**Chronic (n=10)**	**Mixture (n=47)**
Dice (%)	67.05 (19.73)	76.93 (15.33)	80.54 (11.27)	80.66 (8.93)
Recall (%)	73.00 (25.10)	75.59 (25.27)	83.49 (14.70)	82.26 (12.09)
Precision (%)	68.88 (18.16)	81.20 (4.98)	78.60 (11.01)	80.85 (11.20)

In Table 3.2, the standard deviations of both the proposed model and U-net are calculated at the pixel-level and not at the patient-level. Based on Table 3.2, it might be concluded that the overall performance on subjects with acute SDH was lower than the other three types. However, chronic and mixed subdural hematomas tend to be larger in elderly patients because there is more space for the hematoma to expand without symptoms due to brain atrophy. Thus, the proportion of the subjects with acute SDH and less than 25 cc of blood is higher than other types, and, as discussed earlier, these small hematoma regions are more challenging to detect. Thus, in order to have a valid comparison, the subjects were first grouped according to volume, after which the performance was compared for different types. Fig. 3.12 shows the average Dice similarity coefficient for each category. Based on this plot, for moderate and severe SDH, the performance of the proposed segmentation algorithm was consistent among different subdural types. We do not have enough sample size to either reject or accept performance consistency among different types of the mild ($<$ 25 cc) SDH subjects.

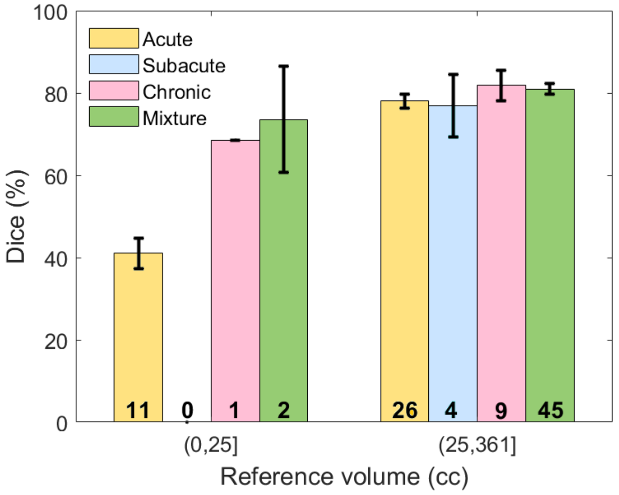


Figure 3.12: The Dice similarity coefficient based on both the severity and type of hematoma. Error bars represent ± 1 SE, the 68% confidence interval. The number of samples for each category is shown on the corresponding bar (*Farzaneh et al.*, 2020).

3.5.4 Inter-physician Variability Analysis

Fig. 3.13 illustrates the comparison between the two neuroradiologists, as well as a comparison of our algorithm's segmentation against the adjudicated ground truth with respect to the selected 20 patients. The average inter-rater Dice similarity coefficient on the selected subset of 20 SDH CT scans is 73.71%, while this coefficient is 50.20% and 77.85%, respectively, when stratifying lesions based on <25 cc and ≥25 cc of blood. On the same subset the algorithm's average Dice similarity coefficient value reached 77.81% for moderate and severe SDHs compared to 77.85% for human raters.

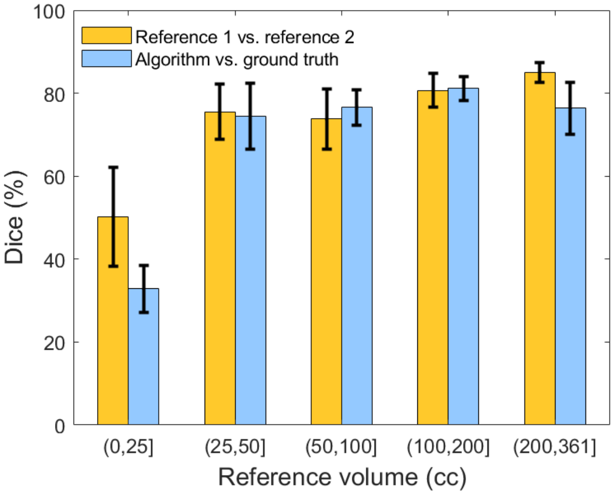
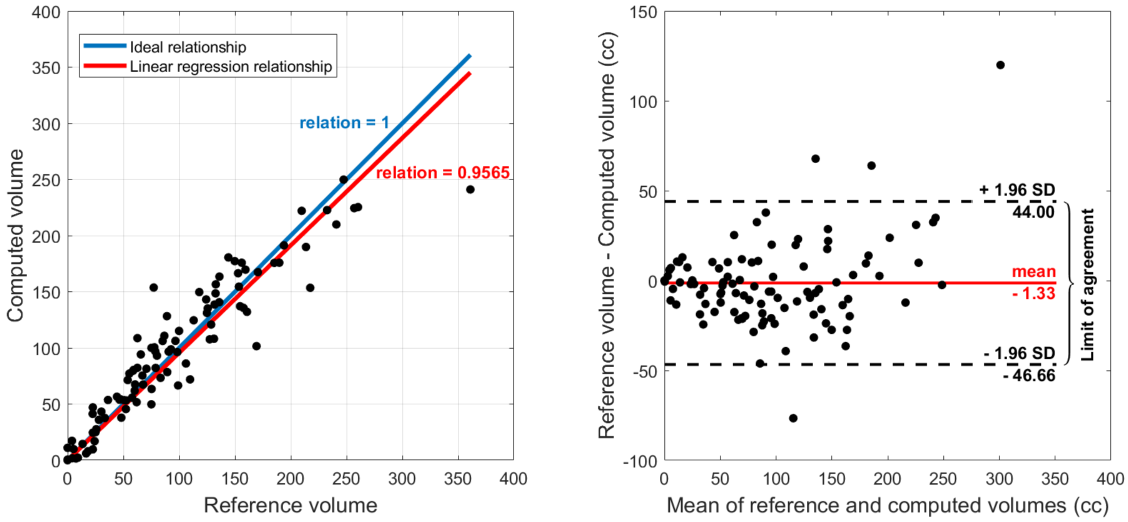


Figure 3.13: Comparison of the Dice similarity coefficient between reference 1, reference 2, and the result of our segmentation algorithm. Error bars represent ± 1 SE, the 68% confidence interval (*Farzaneh et al.*, 2020).

3.5.5 Hematoma Volume Assessment for Severity Analysis

Total hematoma volume was measured for all patients and then used to stratify hematomas by severity. Fig. 3.14a illustrates a linear regression between the volume resulting from the algorithm and the ground truth volume for 110 studied patients. From this plot, it can be concluded that the algorithm tends to under-segment larger lesions while it overestimates the smaller ones. The linear regression relation is 0.96 (SE 0.02, p-value < 0.01).

Next, Bland-Altman analysis was performed to determine the agreement between the reference and computed volumetric measurements. Fig. 3.14b shows the corresponding Bland-Altman plot. This result shows a bias close to zero: -1.33 cc (95% confidence interval -46.44 to 44.00), suggesting that there is no systematic error in calculating SDH volume using the algorithm.



(a) (b)

Figure 3.14: Linear regression relation and Bland-Altman analyses. (a) Linear regression relation between the computed volume and reference volume. Each point corresponds to one patient. (b) Bland-Altman plot that indicates the normality of error (*Farzaneh et al.*, 2020).

One of the application of this analysis was to develop tools that will alert medical providers when a patient requires immediate intervention, which in the context of subdural hematoma corresponds to the ability to identify volumetrically large hematomas. Table 3.3 shows the confusion matrix in classifying patients according to severity based on their lesion volume as defined above. The recall, specificity and F1 score in detecting moderate/severe patients (more than 25 cc of SDH volume) are 98.81%, 92.31%, and 98.22%, respectively.

Table 3.3: Confusion matrix for classifying subdural hematoma severity by volume.

Reference \Calculated	0 - 25 cc	>25 cc
0 - 25 cc	24	2
>25 cc	1	83

3.6 Discussion

By developing a machine learning technique (i.e., a classifier) to divide brain tissue pixels to SDH and non-SDH followed by applying a post-processing step, we performed the SDH segmentation and achieved an average Dice similarity coefficient of 75.35% in segmenting subdural hematoma. The Dice similarity coefficient was 79.97% for moderate and severe SDHs. Our segmentation algorithm reached an average specificity value of 99.89% for healthy subjects, which indicates the algorithm's high performance in avoiding false positive pixels in healthy subjects. The model was shown to be generalizable to different types of SDHs with more than 25 cc of blood. Besides evaluating the method by using Dice, precision, and recall, we performed a Bland-Altman analysis to see if there is any systematic error or bias in the computed volumetric measurement and found neither of these. Next, SDH patients were categorized to severe and non-severe ones and achieved the recall, specificity and F1 score of 98.81%, 92.31%, and 98.22%, respectively, in detecting moderate or severe subjects. However, we should acknowledge that the control cases are from patients with no brain injuries and do not include scans with non-SDH mass lesions. For the population of interest, moderate and severe hematoma cases (i.e., more than 25 cc of SDH), our model is shown to be robust in segmentation of all types of SDH.

In our inter-physician variability study to create a human benchmark, we created two independent sets of ground truth from a representative subset of CT scans. We showed that there is 73.71% agreement between the ground truths created by two skilled neuroradiologists. This finding indicates that there is uncertainty in the gold standard (i.e., manual) labels. This issue is due to the fact that in lesion studies, including hematoma detection, the border between the affected region and the adjacent healthy brain is not necessarily well-separated. Using inaccurate labels to train a machine learning algorithm can adversely affect its performance as well.

To our knowledge, no prior study has previously described automated methods to identify the different SDH types. Given the significant variations in size and intensity of SDHs in CT scans, developing a generalizable algorithm is challenging, but essential because many patients (in particular the elderly) present with a combination of both acute and chronic hematomas. In this work, we sought to address these challenges by proposing a learning algorithm that employs the advantages of both classical image processing and deep learning approaches.

Although the results are comparable with the human benchmark, there are limitations in this study. One of the limitations is the lack of enough samples to investigate the generalizability of the algorithm over different types of blood for mild SDHs. This reflects the pathophysiological mechanism of this type of traumatic injury. Chronic and mixed SDH are more prevalent in the elderly population for whom there is more space inside the skull for the hematoma to expand before experiencing symptoms. Another limitation is that the model is trained on anisotropic resolution images with low resolution retrieved from PACS. PACS down-samples the images along the longitudinal axis while storing them, which can reduce the image resolution. For instance, the pixel spacing along the longitudinal axis after saving them could be over 10 times greater than the spacing along the sagittal and frontal axes (e.g., 5 mm vs less than 0.5 mm). This anisotropic resolution could lead to discontinuity along the longitudinal axis after down-sampling and hence affecting the recall of the algorithm specifically for small hematoma regions.

3.7 Conclusion

Even though quantitative CT characteristics improve diagnosis and outcome prediction, traditional visual examinations — which are qualitative, subjective, and costly — remain the only investigation clinicians perform. Automating the detection and quantitative measurement of moderate and severe SDH could provide a basis to improve diagnostic accuracy and prevent delayed diagnosis. Thus, accurately detecting and quantifying larger and, therefore, more clinically significant subdural hematomas can facilitate timely identification of patients in greatest need of early treatment interventions.

In this chapter, a fully automated approach for segmentation and severity assessment of subdural hematoma is proposed. First, a SDH segmentation model was developed using a combination of deep learning and classical image processing approaches. This model was proven to be robust and generalizable with respect to hematoma type and severity. Next, based on the automatically measured hematoma volume, SDH patients were categorized as severe and non-severe. This model enables the accurate quantification of hematoma severity, which otherwise is almost impossible due to the time-consuming manual process.

CHAPTER IV

Automated Detection and Severity Assessment of Liver Trauma

4.1 Introduction

Approximately 5% of all trauma admissions are attributed to liver trauma (*Taghavi and Askari*, 2019). Due to its anterior location, large size, and fragile parenchyma, the liver is the most frequently injured abdominal organ involved in blunt abdominal trauma (*Ahmed and Vernick*, 2011; *Taghavi and Askari*, 2019; *Badger et al.*, 2009). Early detection and severity assessment of liver trauma with adequate treatment may result in significant reduction of morbidity and mortality (*Piper and Peitzman*, 2010; *Doklestić et al.*, 2015; *Barrie et al.*, 2018).

Contrast-enhanced CT is considered the gold standard technique in evaluating liver trauma and monitoring its progression over time (*Badger et al.*, 2009; *Buquicchio et al.*, 2018). Contrast enhancement is the process whereby the optimal visible difference among adjacent structures (e.g., a lesion and the normal surrounding structure) is obtained by injecting a contrast agent. Depending on the timing of the CT scan capture after initiation of contrast agent injection, abdominal CT contrast phase can be divided into different phases such as arterial, portal venous, nephrographic and delayed (*Birnbaum et al.*, 1996). The CT-driven AAST (American Association for the

Surgery of Trauma) scale is the primary tool currently in use to assess the extent of the liver trauma and guide management (*Croce et al.*, 1991; *Gwinn and Park*, 2020). AAST is a six-point scale with grade I signifying a small subcapsular hematoma (<10% surface area) or laceration (<1 cm parenchymal depth) and grade IV signifying larger laceration with parenchymal disruption affecting 25-75% of either liver lobe or 1-3 liver segments (Couinaud) (*AAST*, Accessed: February, 2021). However, the literature suggests significant intra- and inter-observer variability when visually assessing liver injury using the AAST grading system (*Powers et al.*, 2012; *Nellensteijn et al.*, 2009). In addition to being error-prone, visual examination might be incapable of real-time accurate quantification of the size and severity of abnormalities. Novel big data analytics and computational frameworks, however, are keys to solving such problems in digital health technology (*Choy et al.*, 2018).

One of the main CT imaging criterion in determining AAST grade is the percentage of liver parenchyma that has been disrupted by laceration or intraparenchymal hematoma (*Croce et al.*, 1991). This measurement is referred to as the *liver disruption involvement* (LDI) in the remainder of this chapter. Both liver laceration and intraparenchymal hematoma typically present as regions of low density as compared to adjacent unaffected/normal liver parenchyma. However, the size and shape of liver parenchymal injuries vary significantly depending on the mechanism of injury and severity of the trauma (*Badger et al.*, 2009).

The primary aim of this study is to develop a fully automated image processing and deep learning framework that provides clinicians with quantitative assessment of LDI. This framework can act as a triage tool by rapidly assessing liver injury and its severity. To this end, both the whole liver parenchyma and liver trauma regions are automatically segmented in 3D abdominopelvic CT scans. Accordingly, the percentage of liver parenchyma that is affected by trauma will be computed.

To the best of our knowledge, except for Drezin et al. (*Dreizin et al.*, 2021), no published study has proposed an automated method to segment liver trauma utilizing CT scans. However, since the goal of Drezin et al. is to detect major hepatic artery injury, it focuses on more severe cases and only includes cases with visible liver trauma (no control cases with normal liver) for training and validation purposes. Thus, it takes advantage of prior knowledge that the liver is definitely traumatically injured. Regarding automated liver segmentation, there is a body of literature that has investigated this task, however, all are focused on non-traumatic livers. Those liver segmentation techniques either implement deep learning methods (*Ahmad et al.*, 2019; *Lu et al.*, 2017; *Christ et al.*, 2016) or employ classical image processing techniques (*Farzaneh et al.*, 2016, 2017a; *Lebre et al.*, 2019a,b; *Okada et al.*, 2008). Probabilistic atlases and active shape modeling are among the most popular classical approaches for the liver segmentation task. Farzaneh et al. (*Farzaneh et al.*, 2016, 2017a) proposed a hierarchical approach based on location and customized intensity probabilistic atlases to segment the liver. Lebre et al. (*Lebre et al.*, 2019a,b) used a location probabilistic atlas in combination with shape modeling. Rafiei et al. (*Rafiei et al.*, 2019) also employed a location-based probabilistic model to generate an initial segmentation, which was then refined using an adaptive region growing technique. Okada et al. (*Okada et al.*, 2008) and Shi et. al (*Shi et al.*, 2017) used a probabilistic atlas to generate the initial segmentation mask and then refined it using statistical shape modeling.

The proposed framework enables an objective quantitative assessment of liver trauma as opposed to the sometimes subjective AAST grading system used in current clinical practice. The output of this study can enhance real-time liver trauma diagnostics and be used as a triage tool. Moreover, it can quantitatively measure the volumetric progression or improvement of traumatic injuries at multiple time points, guiding further investigation and management (*Badger et al.*, 2009; *Cuff et al.*, 2000).

4.2 Materials

Before the initiation of this research project, IRB approval (HUM00098656) was obtained. Patient informed consent was not required given that this was a retrospective investigation. This study included 77 patients presented to the University of Michigan Health System (UMHS) Department of Radiology for CT imaging for the evaluation of abdominal blunt force trauma between 01/01/2009 and 8/30/2014, as well as those CTs ordered by the Emergency Department. In total, the 77 CT scans comprised 8072 axial CT slices. Average patient age was 41.43 years, with a range of 18-88 years. Of the 77 patients included in this investigation, 34 had evidence of liver trauma and 43 had no evidence of liver parenchymal disruption on contrast-enhanced CT scan.

All CT scans were acquired in the axial plane using either GE Medical Systems (LightSpeed VCT or Discovery CT750 HD models) or SIEMENS (Emotion 16 model). Trauma protocol CT scans often include both an arterial and portal venous phase to evaluate for both arterial (e.g., aortic) and solid organ injury. Only the portal venous phase was utilized in this study as this phase is optimal for the detection of hepatic parenchymal injuries.

To generate ground truth for all 77 patients, livers were manually annotated, which meant that the margins of the liver itself were outlined. Next, any liver laceration or hematomas were manually annotated for 34 CT scans with visible liver parenchymal disruption. CT scans. Each CT scan was manually annotated slice by slice to generate binary masks (i.e., ground truth) for injury and organ. All annotations were verified by a fellowship-trained abdominal radiologist with 5 years of post-training experience.

4.3 Methods

4.3.1 Liver trauma assessment framework

The study design for liver trauma segmentation and severity assessment is shown in Fig. 4.1. First, deep learning-based models are developed to segment both liver organ and trauma regions. Then, to assess the severity of the liver trauma, the automatically segmented regions are processed to measure liver disruption volume and, accordingly, calculate the proportion of the liver tissue affected by those injuries (i.e., LDI).

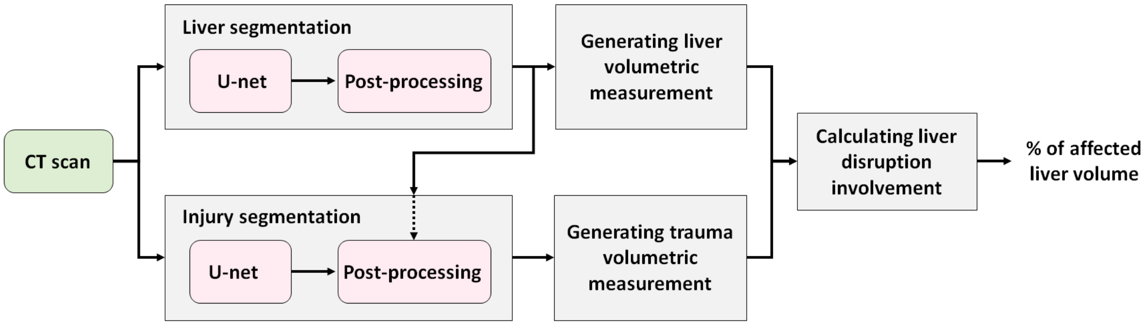

CT scan
Liver segmentation
U-net
Post-processing
Injury segmentation
U-net
Post-processing
Generating liver volumetric measurement
Generating trauma volumetric measurement
Calculating liver disruption involvement
% of affected liver volume

Figure 4.1: A high-level study design for liver trauma segmentation and severity assessment (*Farzaneh et al.*, 2021a).

.

4.3.2 Liver segmentation

Fig. 4.2 demonstrates a high-level overview of the proposed liver segmentation method.

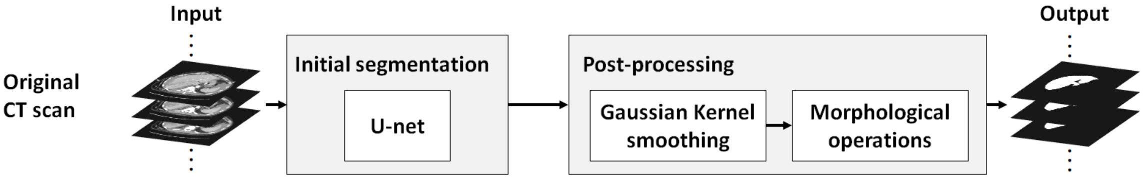


Figure 4.2: A schematic diagram of the proposed liver segmentation method (*Farzaneh et al.*, 2021a).

With a contrast-enhanced CT scan, we first employed a U-net model (*Ronneberger* *et al.*, 2015) to generate the initial liver mask. In the proposed model, data augmentation was performed by rotating, re-scaling, and translating the images to enhance the training data set. Next, the post-processing module transformed the volumetric masks from the U-net model into the final segmentation map. To that end, the initial mask was filtered using 3D Gaussian kernel smoothing to achieve spatial coherency and smooth binary mask contours according to the neighboring pixels. Finally, morphological operations were used to remove small, sparse regions; fill the holes in axial planes; and exclude any region that was not connected to the largest 3D connected component.

4.3.3 Liver disruption segmentation

As shown in Fig. 4.3, a second U-net backbone model was trained to segment the liver trauma regions. The post-processing module comprises the volumetric reconstruction of the U-net output, during which human domain knowledge regarding the location and intensity distribution of liver trauma was integrated into the model. Considering that trauma regions are within the liver parenchyma, if more than 50% of the initial segmented trauma mask fell outside the segmented liver, the region would be excluded.

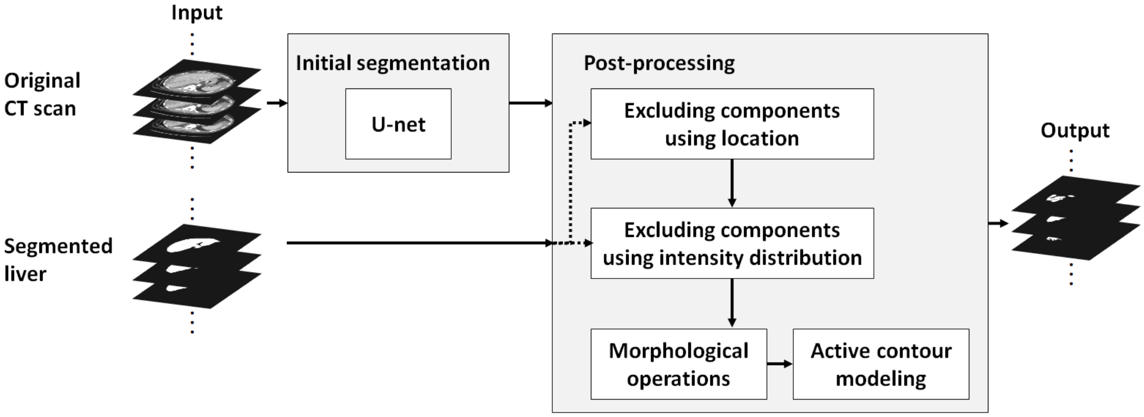


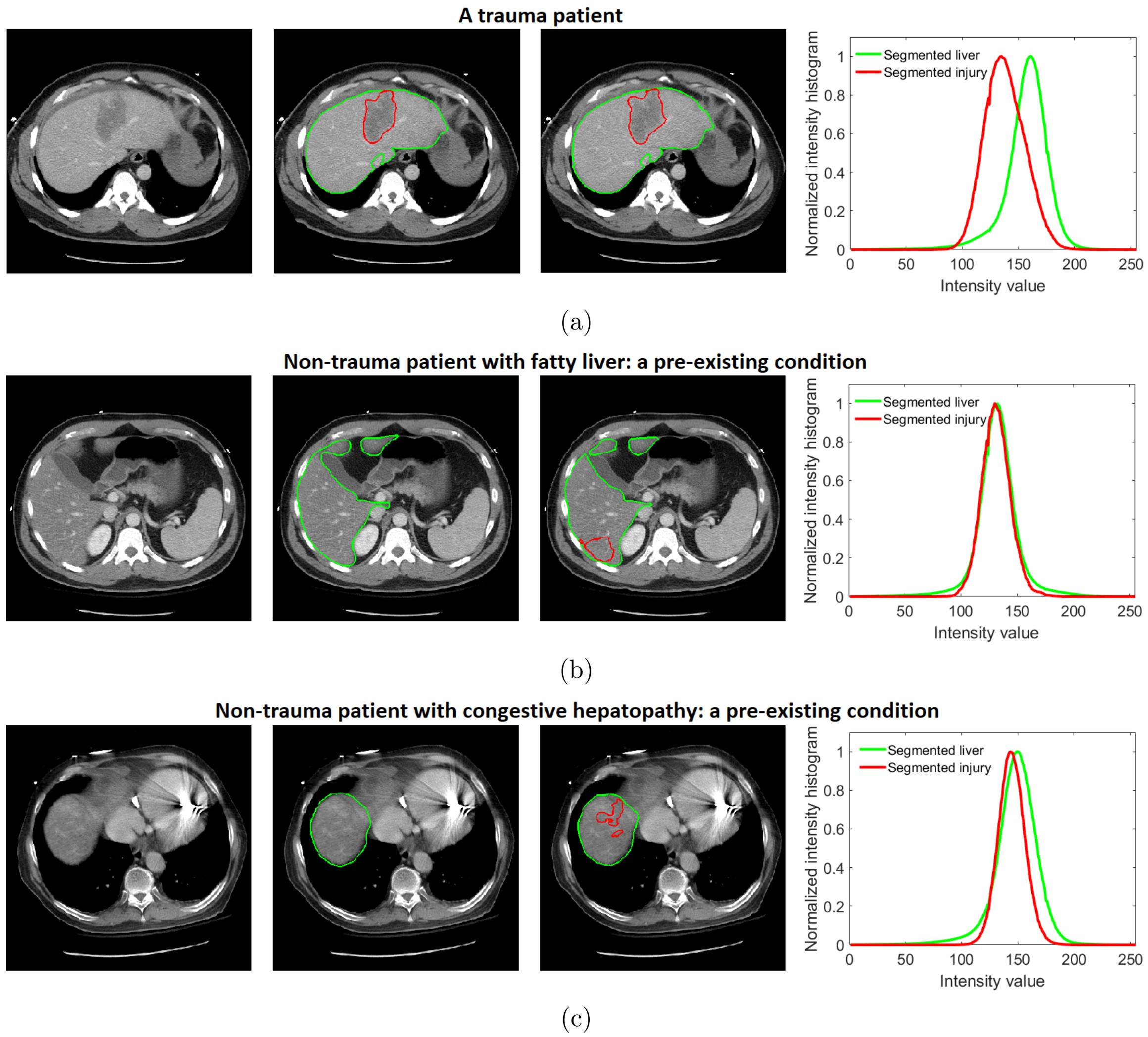
Figure 4.3: A schematic diagram of the proposed liver trauma segmentation method (*Farzaneh et al.*, 2021a).

Pre-existing conditions such as fatty livers (Fig. 4.4b) or congestive hepatopathy (Fig. 4.4c) lead to the different representation of non-trauma liver parenchyma on CT scans (*Bydder et al.*, 1981; *Sass et al.*, 2005; *Wells et al.*, 2016). In theory, these pre-existing conditions could cause the U-net model to falsely detect trauma given the presence of low-attenuation of the parenchyma at baseline (Fig. 4.4a). To exclude regions falsely segmented as trauma (e.g., part of the normal liver parenchymal), two intensity distributions were generated, corresponding to: 1) pixels of the CT image segmented as the liver, and 2) pixels of the CT image segmented as liver trauma (Fig. 4.4). Next, the means of these two distributions were compared using a two-sample t-test. If the test statistic value was less than a fixed threshold, we concluded that the two intensity distributions were from the same texture and thus the segmented trauma region was part of the non-trauma liver parenchyma. Correspondingly, these false positive components were excluded from the segmentation using the intensity distribution.

Next, the 3D Chan-Vese Active Contour Model (ACM) (*Chan and Vese*, 2001) was used to iteratively evolve the boundary of the initial segmentation according to local intensity and spatial coherence. The energy function $F(s_1, s_2, S)$ was defined as

$$\begin{aligned} F(s_1, s_2, S) &= \mu \cdot A(S) + \nu \cdot V(S) \\ &\quad + \lambda_1 \int\limits_{\text{inside}(S)} |I(x,y,z) - s_1|^2 \, dx \; dy \; dz \\ &\quad + \lambda_2 \int\limits_{\text{outside}(S)} |I(x,y,z) - s_2|^2 \, dx \; dy \; dz, \end{aligned} \tag{4.1}$$

where S is the current surface, and s_1 and s_2 respectively correspond to the average intensities inside and outside the surface S. $I(x, y, z)$ denotes the intensity value of a pixel at the (x, y, z) coordinate. Moreover, $A(.)$ and $V(.)$ calculate the area and volume of a surface respectively. In Equation (4.1), parameters μ, ν, λ_1, and λ_2 are



A trauma patient

(a)

Non-trauma patient with fatty liver: a pre-existing condition

(b)

Non-trauma patient with congestive hepatopathy: a pre-existing condition

(c)

Figure 4.4: Liver trauma and organ segmentation results as well as ground truth annotations in patients with trauma or pre-existing conditions. (a) CT image shows an intraparenchymal hematoma in the left liver lobe. (b) CT image shows diffuse low attenuation of the the right liver lobe relative to the spleen, consistent with fat deposition. This is a non-traumatic pre-existing condition. (c) CT image shows heterogeneous enhancement of the right liver dome due to congestive hepatopathy, a non-traumatic pre-existing condition. In (a), (b), and (c), the first column corresponds to the original CT image; the second column corresponds to the ground truth annotations; and the third column corresponds to the automated segmentation results before post-processing. In both ground truth annotations and segmentation results, the green line shows the liver contour while the red line marks the contour of trauma regions. The fourth column compares the pixel intensity distribution inside the segmented liver and the segmented trauma region. As the t-test indicated the difference between the two means of the aforementioned distributions was small for both examples of (b) and (c), the corresponding segmented injured regions were excluded during the post-processing step (*Farzaneh et al.*, 2021a).

constants. Following the Chan-Vese paper, parameters λ_1 and λ_2 were set to 1. The parameter μ, which specifies the degree of smoothness of the segmented region, was set to 0.1 based on prior work on an independent medical image processing problem (*Farzaneh et al.*, 2018). Finally, the parameter ν controls contraction bias, which specifies the tendency of the active contour to grow outward. This parameter was determined using a grid search. To evolve the contour, at each iteration a Sparse-Field level-set method, similar to the one proposed in Whitaker et al. (*Whitaker*, 1998) was implemented. After each iteration, the mask was modified to exclude the added pixels that fell outside the automatically segmented liver. Finally, morphological operations were applied to remove small, sparse regions and fill the holes in the axial plane. The effect of the post-processing step is analyzed in the result section.

4.3.4 Liver disruption involvement measurement

After segmenting both liver and trauma regions (when present) as binary masks, the volumes were estimated according to respective pixel maps and the unit pixel volumes (i.e., number of pixels from the binary mask × unit pixel volume). The unit pixel volume was calculated using slice spacing and pixel spacing values extracted from CT scan metadata.

LDI is then estimated as

$$\text{LDI}(\%) = \frac{\hat{V}(trauma)}{\hat{V}(liver)} \times 100, \tag{4.2}$$

where $\hat{V}(.)$ corresponds to the estimated volume of a segmented region. For patients with no detected traumatic injury to the liver, the trauma region volume was set to zero.

4.3.5 Statistical Analysis

A comprehensive evaluation of the segmentation model's performance was performed on the validation sets using Dice similarity coefficient, recall, precision, Relative Volume Difference (RVD), and Volumetric Overlap Error (VOE). RVD and VOE error measures were calculated according to definition from Heimann et al. (*Heimann et al.*, 2009) as

$$\begin{aligned} \text{RVD} &= \frac{|S| - |GT|}{|GT|} \\ \text{VOE} &= 1 - \frac{|S \cap GT|}{|S \cup GT|}, \end{aligned} \tag{4.3}$$

where GT and S correspond to the ground truth and segmented masks, respectively, while $|\,.\,|$ computes the number of pixels in the corresponding mask.

To measure the variability in the LDI estimates, linear regression analysis was performed in which the computed and reference LDI measurements were plotted against each other. The linear regression relation between the two measures was then calculated. Moreover, to better perceive the algorithm's agreement with the ground truth and potential systematic errors, a Bland-Altman analysis was employed (*Bland and Altman*, 1986, 2010).

4.4 Results

For this investigation, the presence and extent of liver trauma were assessed using the percentage of the liver affected by traumatic injuries. To compute the percentage of liver disruption, both liver and trauma regions were segmented using deep learning and image processing techniques.

This is a secondary study of an internal data set from the UMHS that includes 77 patients, among whom 34 experienced trauma-related liver parenchymal disruption and 43 had no evidence of liver parenchymal disruption at imaging. To train and validate the segmentation models, patient-wise five-fold cross-validation was implemented. Folds were created to include a roughly balanced distribution of the trauma severity in terms of the reference LDI. The cross-validation folds remained the same for both liver and trauma segmentation tasks.

4.4.1 Liver segmentation

The performance of the proposed liver segmentation algorithm is shown in Table 4.1. Our algorithm yielded mean Dice, recall, and precision values of 96.13%, 96.00%, and 96.35%, respectively, when tested on the internal UMHS data set. In order to evaluate our segmentation model, in addition to the UMHS data set, we used the publicly available 3DIRCAD data set that includes 20 pathological CT scans with hepatic tumors in 75% of the cases (*3DIRCAD*, Accessed: February, 2021). Although the imaging parameters and underlying pathology of the internal UMHS and 3DIRCAD data sets are different, the 3DIRCAD data set was only employed for testing purposes using the U-net trained only on the UMHS data set. Based on Table 4.1, it can be concluded that the overall performance of the proposed algorithm is comparable with the state-of-the-art models even without tuning the weights of U-net, which indicates the generalizability of the proposed model on an unobserved data set.

Fig. 4.5 shows the average Dice similarity score stratified by severity of trauma. As shown, the performance on severe cases with more than 20% of liver disruption is slightly lower as compared to smaller injuries (average Dice = 94.14%); this could be due to extensive injuries distorting the contour of the liver itself.

Table 4.1: Performance of our proposed liver segmentation approach compared with state-of-the-art methods. Numbers in parentheses are standard deviations. For the cited studies, scores are reported as presented in the original papers. Ahmad et al. used a different definition of VOE. NR: not reported.

Method	**Data set**	**Dice** (%)	**Recall** (%)	**Precision** (%)	**RVD** (%)	**VOE** (%)
Proposed Method	UMHS Data set (n=77)	96.13 (1.49)	96.00 (2.83)	96.35 (2.09)	-0.30 (4.24)	7.40 (2.69)
Proposed Method	3DIRCAD (n=20)	94.64 (2.18)	95.06 (4.07)	94.38 (2.75)	0.83 (5.79)	10.10 (3.85)
(*Ahmad et al.*, 2019)	Subset of 3DIRCAD (n=5)	91.83 (1.37)	NR	NR	5.59 (6.49)	6.09 (1.21)
(*Lu et al.*, 2017)	3DIRCAD (n=20)	NR	NR	NR	0.97 (3.26)	9.36 (3.34)
(*Christ et al.*, 2016)	3DIRCAD (n=20)	94.3	NR	NR	-1.4	10.7
(*Lebre et al.*, 2019a)	3DIRCAD (n=20)	88 (3)	87 (5)	89 (4)	NR	NR

4.4.2 Liver disruption segmentation

Table 4.2 and Fig. 4.6a compare the liver trauma segmentation results with the ground truth. To investigate the generalizability of the algorithm with respect to injury severity, the results are stratified based on the reference LDI level. These results show that the post-processing step improves the performance in terms of the Dice similarity score. The final model achieved an average Dice of 51.21% in segmenting liver disruption while this value reached 72.45% for considerable liver disruptions that involved more than 5% of the liver. Possible etiologies for the lower performance measurements for smaller injuries with less than 2% of LDI include: 1) those subtle injuries are inherently more challenging to segment, and 2) for smaller regions, even small deviations have a greater adverse impact on the performance metrics.

Since the performance metrics including Dice, recall, and precision are not de-

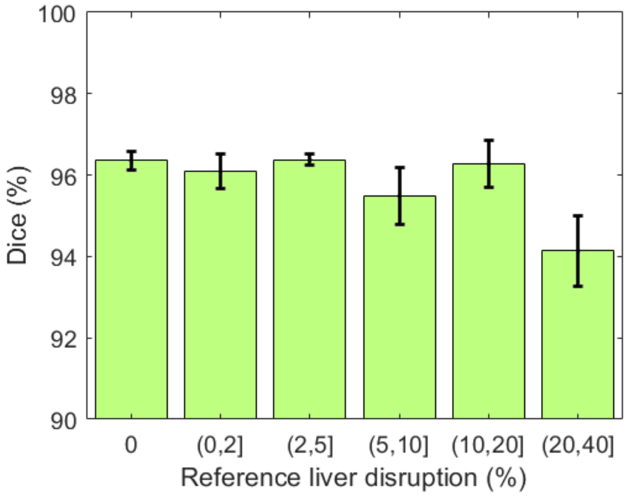


Figure 4.5: The Dice similarity coefficient for liver segmentation stratified based on the liver disruption involvement. Error bars represent ± 1 standard errors, the 68% confidence interval (*Farzaneh et al.*, 2021a).

Table 4.2: Performance of our proposed liver trauma segmentation approach stratified based on the severity of the injury. Numbers in parentheses are standard deviations.

Method	**% Liver Disruption**	**Dice (%)**	**Recall (%)**	**Precision (%)**	**RVD (%)**	**VOE (%)**
Proposed Method	0 – 2 % (n=15)	28.06 (23.09)	32.15 (32.90)	35.66 (27.01)	15.83 (113.41)	81.68 (16.01)
Proposed Method	2 – 5 % (n=4)	58.35 (17.91)	54.78 (25.39)	66.21 (6.93)	-19.16 (30.85)	57.05 (18.64)
Proposed Method	>5 % (n=15)	72.45 (11.82)	73.84 (18.60)	75.35 (11.51)	1.86 (36.03)	42.02 (13.64)
Proposed Method	All (n=34)	51.21 (27.74)	53.20 (32.56)	56.76 (27.21)	5.55 (78.88)	61.29 (24.07)
U-net (No Post-processing)	All (n=34)	47.75 (27.61)	47.32 (31.13)	56.64 (28.97)	-2.71 (63.10)	64.50 (23.80)

fined for the cases without any liver trauma, to evaluate the non-trauma cases, the computed LDI was used, which, ideally, should be zero for non-trauma patients. Fig. 4.6b compares this value for non-trauma cases before and after applying the post-processing step. The average LDI for these cases is 0.27% after post-processing, with none over 2.6%. This close to zero performance shows the accuracy of the algorithm in avoiding false positives. The patient marked by "Patient X" in Fig. 4.6b is the patient with congestive hepatopathy, a pre-existing condition shown in Fig. 4.4c. Fig. 4.4c shows the falsely segmented region before post-processing. This false positive region was excluded through the post-processing step with respect to the customized intensity distribution of the intact liver. As a result, the computed LDI of Patient X is reduced to 1.61% from 8.27%.

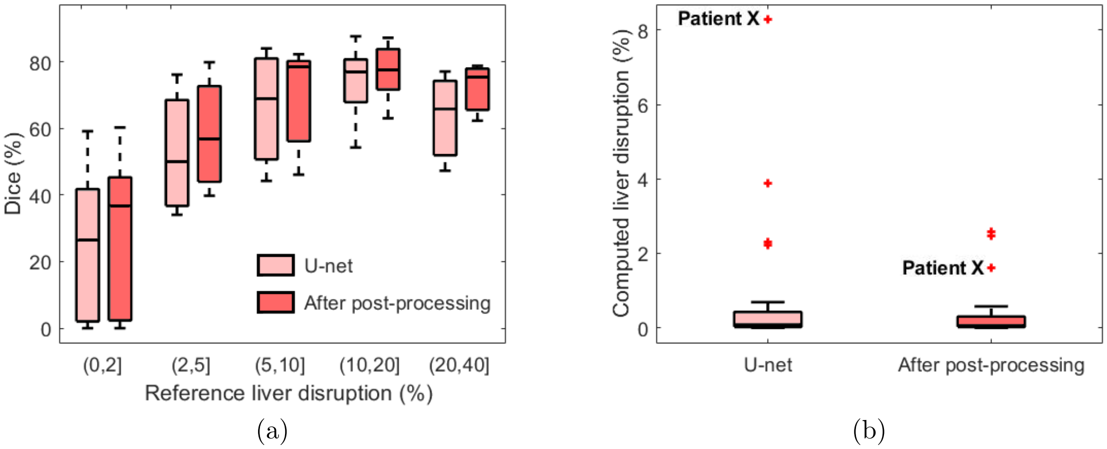
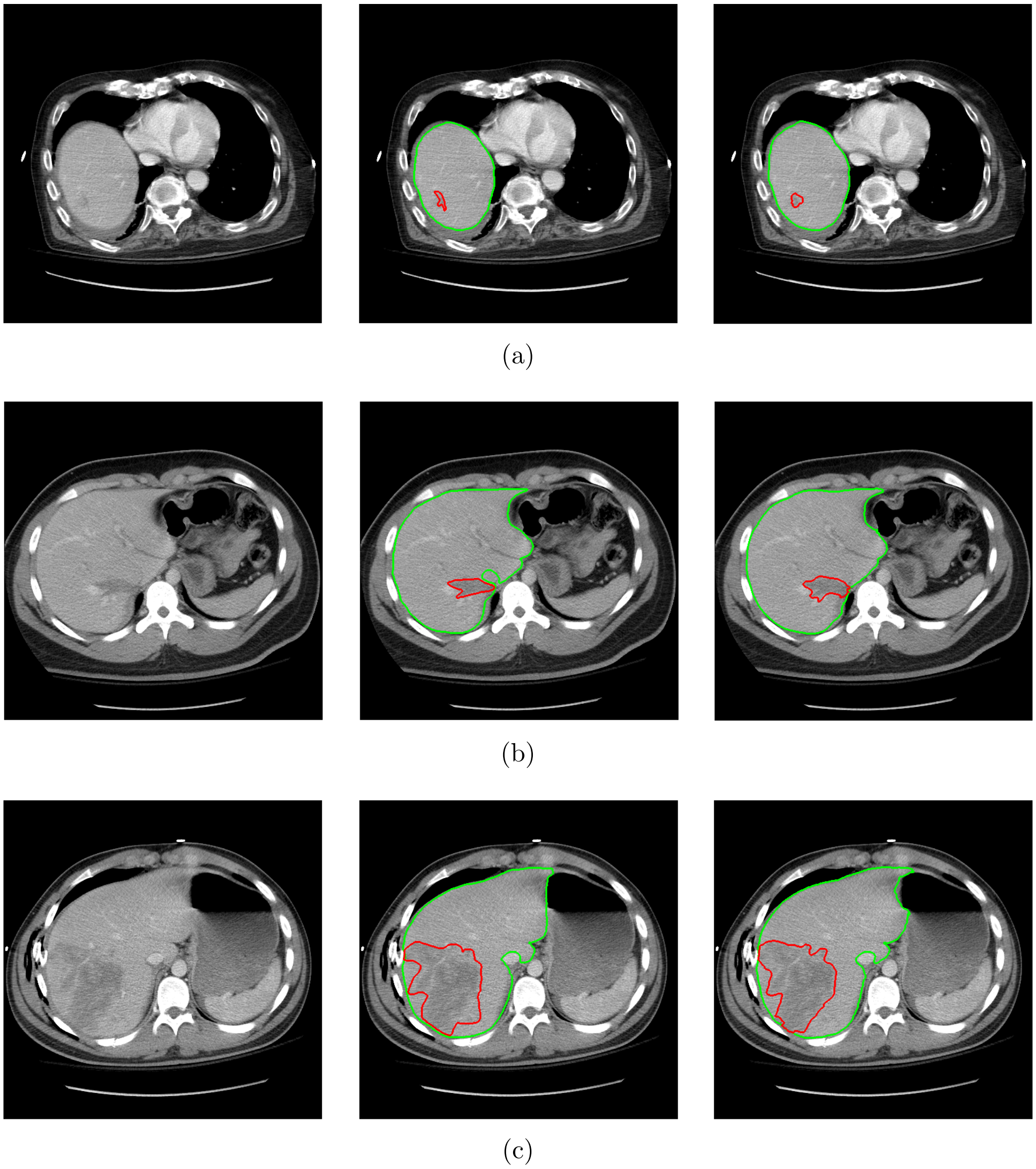


Figure 4.6: (a) A box plot comparing the Dice similarity coefficients of the proposed injury segmentation algorithm and U-net with respect to the percentage of liver disruption. (b) A box plot comparing the computed LDI for cases without liver trauma. The reference LDI is zero (*Farzaneh et al.*, 2021a).

The results of the proposed liver and trauma segmentation approaches are shown in Fig. 4.7. The images cover various severity levels of liver trauma. These results demonstrate that the proposed deep learning-based framework can accurately assess liver trauma, a heterogeneous clinical problem, in an automated and quantitative fashion.

4.4.3 Liver disruption involvement (LDI) measurement

LDI measures the percentage of the liver parenchyma affected by blunt traumatic injuries. Fig. 4.8a compares the computed versus reference LDI measures for all 77 studied patients. The linear regression relation is 0.95 with p-value < 0.01.

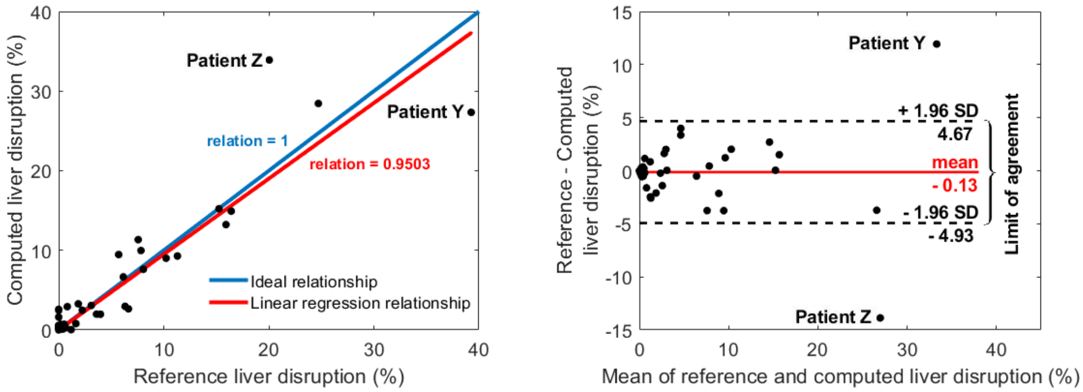

(a)

(b)

(c)

Figure 4.7: Liver trauma and organ segmentation results along with ground truth annotations in three separate patients with different levels of LDI. (a) a CT image from a patient with a very subtle 0.62% liver disruption. (b) a CT image from a patient with 2.25% reference LDI. (c) a CT image from a patient with 15.26% reference LDI. In (a), (b), and (c), the first column corresponds to the original CT image; the second column corresponds to the ground truth annotations; and the third column corresponds to the automated segmentation results. In both ground truth annotations and the segmentation results, the green line shows the liver contour while the red line marks the contour of trauma regions (*Farzaneh et al.*, 2021a).

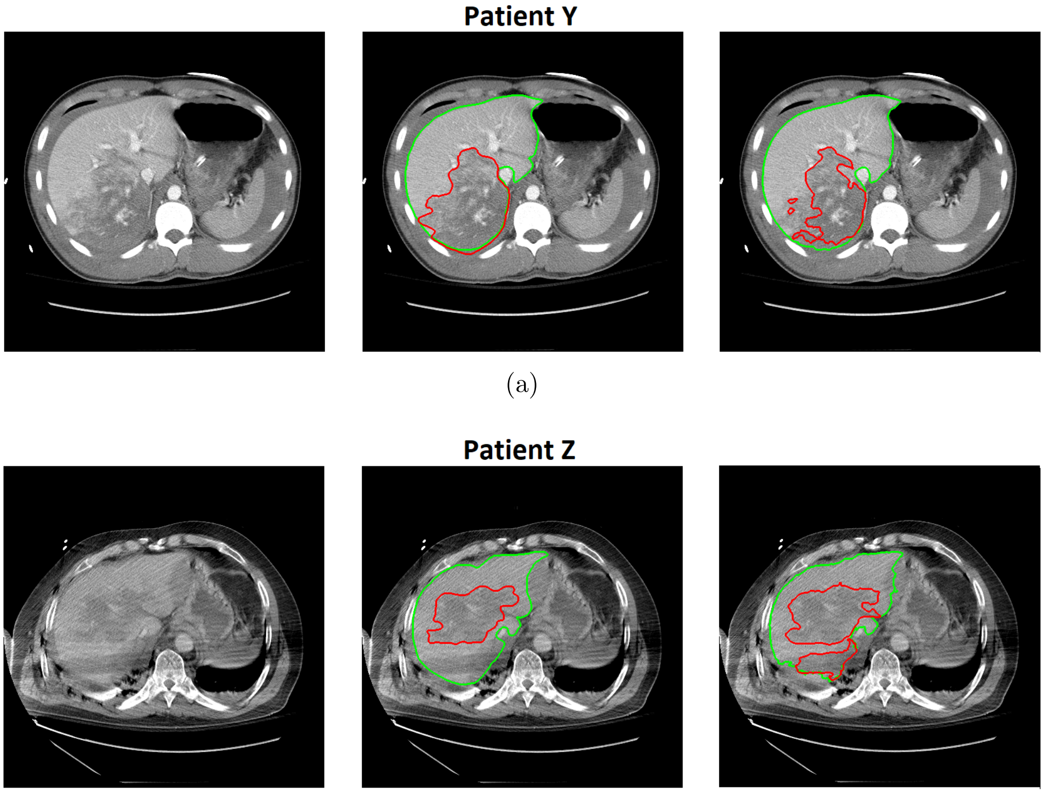
Bland-Altman analysis was also performed to understand the potential systematic errors in computing the hepatic disruption involvement. It can be concluded from



(a) (b)

Figure 4.8: Linear regression relation and Bland-Altman analyses. (a) Linear regression relation between the computed and reference LDIs. Each point corresponds to one patient. (b) Bland-Altman plot that indicates the normality of error (*Farzaneh et al.*, 2021a).

Fig. 4.8b that there is a negligible bias (-0.13%) with 95% confidence interval of -4.93 to 4.67%. Moreover, the Bland-Altman plot indicates two outliers that are marked as “Patient Y” and “Patient Z” in both Fig. 4.8a and 4.8b. Patient Y (Fig. 4.9a) presented with a massive liver disruption affecting 40% of the liver parenchymal, which is over 1.6 times greater than the next largest liver disruption. Since the model had not seen such a large injury in the training phase, it can be concluded that it failed to learn such a pattern, and did not segment the whole injured region. Patient Y’s recall and precision scores are, respectively, 63.68% and 92.24%, indicating the algorithm’s high performance in avoiding false positives. Patient Z’s CT scan shows a strong beam hardening artifact (*Lebre et al.*, 2019a). This streaking artifact appeared as a dark band and was misdiagnosed by the algorithm. This error might have occurred as there were no similar cases in the training folds. These issues can potentially be addressed by extending the data set in the future.



(a)

(b)

Figure 4.9: Liver trauma and parenchymal segmentation results on two patients who were determined to be outliers based on Bland-Altman analysis. (a) a CT image from Patient Y in Fig. 4.8. Patient Y has the largest trauma region in our dataset. (b) a CT image from Patient Z in Fig. 4.8. Patient Z's CT image is distorted by linear streak artifact, which leads to a large false positive region of segmented trauma. In (a) and (b), the first column corresponds to the original CT image; the second column corresponds to the ground truth annotations; and the third column corresponds to the automated segmentation results. In both ground truth annotations and the segmentation results, the green line shows the liver contour while the red line marks the contour of trauma regions (*Farzaneh et al.*, 2021a).

4.5 Discussion

The purpose of this study is to develop an end-to-end framework that can detect and quantitatively assess the severity of liver traumas with respect to the percentage of liver parenchyma injured. For this purpose, the percentage of the liver parenchymal affected by traumatic injuries was automatically computed, as it is one of the main measurements considered in the AAST liver injury scale. The proposed framework provides real-time quantitative information about the injury that was not accessible before due to the cumbersome manual process to annotate all images included in a 3D CT scan. As a result, we envision that this system enables objective, continuous injury severity scoring in the future as an alternative to the current AAST grading. Moreover, this system can be used as a triage tool by rapidly assessing liver injury and its severity as well as for monitoring volumetric progression or improvements of the trauma region at multiple time points.

The proposed algorithm employed a deep learning backbone to segment the initial liver parenchyma and trauma masks. These masks were then refined during a post-processing step by integrating human domain knowledge about the location and intensity of injury into the model. The model achieved Dice similarity coefficients of 96.13% and 51.21%, respectively, in segmenting liver and trauma regions. The Dice score for liver trauma segmentation reached 72.45% for considerable injuries with over 5% of LDI. Moreover, of the 43 non-trauma cases, 40 patients were detected to have <1% of LDI showing high performance of the model in avoiding false positives. With regard to creating the diagnostic model, our algorithm achieved a linear regression relation of 0.95 between the computed versus reference LDI measurements. It can be concluded that the proposed algorithm can accurately quantify the extent of liver parenchymal trauma.

To our knowledge, no prior study has previously described automated methods to identify and assess the severity of liver trauma using abdominopelvic CT images. The mechanism and severity of trauma can lead to significant variations in the size and shape of injured regions on CT scans. Moreover, non-traumatic pre-existing conditions, such as fatty liver and congestive hepatopathy, may significantly affect the liver parenchymal's attenuation in CT scans. Given these sources of variation, developing a generalizable algorithm is challenging but necessary. In this work, we sought to address these challenges by integrating human domain knowledge about the location and intensity distribution of injuries into the model during the post-processing step. We also took advantage of classical image processing techniques, including 3D active contour modeling, to bring spatial coherency and intensity homogeneity to the segmented region.

The present study has a few limitations. First, despite a comprehensive process to generate the ground truth, the reference labels used to train and evaluate the models are best estimates rather than a definitive label. This is because the edges of the injury and organs are not always distinctly visible in CT images. This issue introduces not only noise into the training phase but also uncertainty into the labels against which the performance is measured that can adversely affect the performance metrics. For example, as shown in Fig. 4.7a and 4.7c, although the ground truth and automated segmentation results for trauma do not thoroughly overlap, the segmented regions do not visually appear to be less accurate than the ground truth. The adverse effect of these inconsequential deviations on the performance metrics is greater on smaller regions.

In addition to imperfect labels, artifacts introduced into CT scans during image collection (Fig. 4.9b) are another source of noise. In this study, the CT scans affected by artifacts are not excluded as long as radiologists can make a diagnosis. While the results on CT scans with strong artifacts are not perfect, excluding those scans from the study would make the study less representative of routine clinical practice. Another limitation involves the lack of enough massively injured cases (Fig. 4.9a) in our data set to effectively learn their patterns. The low representation of massive cases could be because those patients went straight to surgery and did not have a CT scan prior to intervention.

4.6 Conclusion

This chapter focused on the development of a framework to assess the severity of liver parenchymal trauma. The model employs a deep learning backbone to generate initial liver parenchyma and trauma masks in a 3D fashion. The initial mask is then refined with respect to the domain knowledge regarding the location and intensity distribution of intact parenchyma and trauma. Next, the severity of liver trauma is quantified as the proportion of the liver parenchyma that is disrupted by the injury. This model is generalizable to heterogeneous-appearing livers in CT scans of patients with pre-existing liver conditions, including fatty liver and congestive hepatopathy. The accuracy of the model for both blunt trauma and non-trauma patients supports this system's potential for enhancing the medical decision-making process.

CHAPTER V

Automated Kidney Segmentation for Patients with Abdominal Trauma

5.1 Introduction

Automated decision support systems could help physicians in clinical diagnosis, prognosis and ultimately treatment planning. Since the main inputs of such systems is images, a major component of these systems is image processing, and in particular image segmentation. Anatomical structures and other regions of interest are delineated using image segmentation techniques. However, due to biological variations, existing noise and artifacts that are inherent components of medical images, grayscale similarity between the border of an organ and neighboring tissue, and different scanner settings, medical segmentation is a challenging task. In the past three decades, many segmentation techniques have been proposed for different medical applications such as anatomical structure segmentation (*Lin et al.*, 2006; *Skalski et al.*, 2017; *Khalifa et al.*, 2016, 2017; *Wolz et al.*, 2012, 2013) and pathology detection (*Liu et al.*, 2015; *Soler et al.*, 2001; *Park et al.*, 2001; *Vorontsov et al.*, 2017). The aims of these methods are to improve upon conventional user-guided segmentation methods and also to delineate images for further quantitative analysis (e.g., volumetric measurement) (*Lin et al.*, 2006), which provides important information for diagnosis/prognosis. The availability of such information in real-time for trauma injury patients is of utmost importance as time is crucial to outcome. Therefore, developing fast and accurate computational methods to segment the organs and detect injuries is of great value.

In this chapter, we focused on trauma injury patients and designed an automated method to segment the kidneys from CT scans. Contrast-enhanced CT scans is the modality of choice to detect kidney injuries. Each contrast phase highlights certain types of kidney trauma and thus contrast phase knowledge leads to more accurate CT image segmentation and lesion classification (*Kawashima et al.*, 2001; *Dayal et al.*, 2013). However, this information is often missing or inaccurate (*Sofka et al.*, 2011). Therefore, it is important for an algorithm to be generalizable over all types to be applicable in real clinical settings.

A number of automated methods have been proposed to segment the kidneys using CT scans. Lin et al. (*Lin et al.*, 2006) categorized kidney segmentation methods to 1) thresholding and region-based; 2) knowledge-based; and 3) deformable methods. They also designed a method using the geometric location of kidney to detect seed points as well as adaptive region growing to segment the kidneys. This method was based upon the assumption that kidney is visible in the middle slice of the set of abdominal CT scan, which might not be always true. Skalski et al. presented a kidney segmentation model that falls in the category of deformable models (*Skalski et al.*, 2017). In this work, ellipsoid shape constraints were incorporated into the level-set formulation. In addition to three kidney segmentation categories introduced in (*Lin et al.*, 2006), over the last decade a body of literature has focused on machine learning approaches such as deep learning and ensemble learning (such as random forest) methods to address this problem. Khalifa et al. (*Khalifa et al.*, 2016, 2017) designed a 3D kidney segmentation algorithm using a random forest classifier and adaptive shape modeling. Wolz et al. (*Wolz et al.*, 2012, 2013) introduced a hierarchical subject-specific atlas generation model to address high inter-subject variability. This model required a large training data set to be practical. Additionally, all of the CT scans were captured at the portal venous contrast phase; therefore, although this model showed promising results its accuracy and generalizability over heterogeneous cohorts in different clinical settings are questionable.

In this chapter, we proposed an automated kidney segmentation method by implementing machine learning and active contour modeling. This algorithm was designed based on CT scans without prior knowledge of contrast phase. Moreover, our focus was on patients admitted to the trauma service which adds to heterogeneity in our medical data.

5.2 Materials

In total, 1750 axial CT images from 35 patients were studied. All CT scans were collected from patients who were admitted to the trauma service at UMHS. Among 35 patients, 22 were transferred to the ICU, 7 of them to a general admission and non-specialty unit bed, 5 to the operating room and 1 patient died. All images were collected using the GE Medical System scanner and shared the same slice thickness of 5 mm, however, contrast phases were different and unknown. The ground truth kidney annotations were verified by an expert radiologist. It is noteworthy that generating the ground truth labels is time consuming and so never performed in clinical settings.

5.3 Methods

5.3.1 Kidney Segmentation Framework

The schematic diagram of the proposed method is depicted in Fig. 5.1. In order to detect the kidneys, we first localized and aligned the abdominal cavity. Towards this end, we adjusted the image contrast, segmented the main bones' mask, and accordingly registered each 3D CT image set using orientation, scaling and transformation functions. Next, a 3D initialization mask was be detected within the abdominal cavity. For this purpose, we divided the abdominal cavity to small patches using a superpixel algorithm, and for each patch we extracted multiple features. These features were then used in our random forest classifier to detect potential initialization pixels. Finally, an adaptive contour model was designed to evolve the boundary of the mask. This method was performed twice to segment the left and the right kidneys independently, after which the results from each execution were combined to produce the final segmentation.

5.3.2 Abdominal Cavity Alignment

CT scans from different patients did not necessarily have the same abdominal size in terms of the diameter and length in transverse plane. Moreover, abdominal regions were not all placed at the same orientation and location in a 2D axial image. Thus, in order to extract features for kidney segmentation, images should be normalized to have the same location, size and rotation angle. For this purpose, first, a CT image was selected as a reference image. Having a new set of slices, the contour of abdominal region was calculated and images were oriented, translated, and scaled to have the same rotation angle, center, and size as the reference slice.

5.3.2.1 Abdominal Contour Calculation

In order to make the mentioned transformation, we first need a criterion to define abdominal area in an image. We used bones such as ribs, sternum, and vertebral column to find the abdominal area. Therefore, the first step was to detect bones in an image.

In each slice, bones were enhanced using linear contrast adjusting based on Hounsfield unit and then segmented using a thresholding method. In each slice, since abdom-

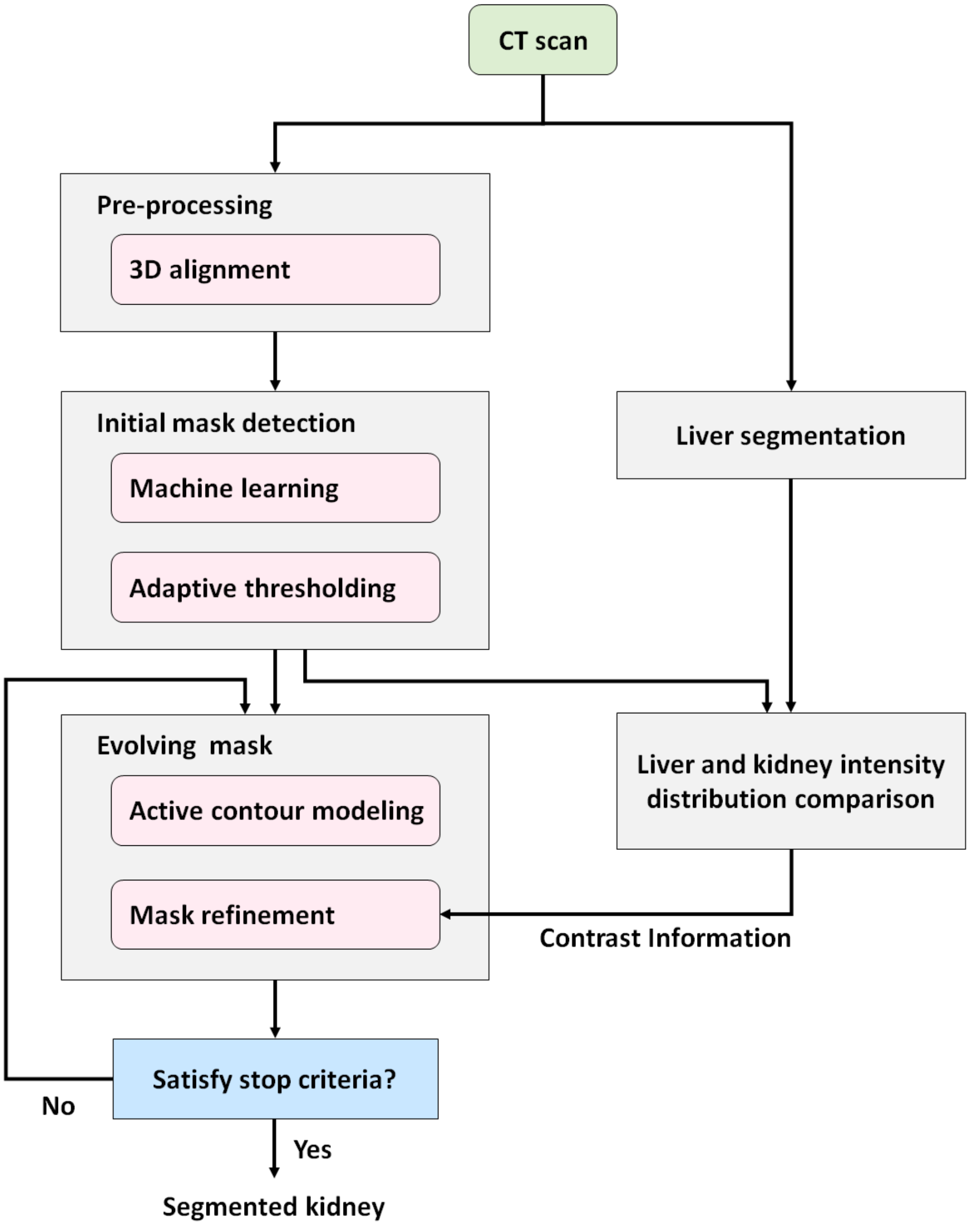


Figure 5.1: Schematic diagram of the proposed kidney segmentation method.

inal bones were not all connected, they barely represented the contour around the abdominal region. Therefore, estimation of the whole bone contour was needed. To estimate the contour, bones in the first top N slices were selected and all the bones were stacked (i.e., superimposed) in one image. In other words, if I_1, ...,I_N show the top N slices after bone segmentation, we created the image $I_s = I_1|I_2|...|I_N$ where '|' shows the logical *or* operator. After stacking, the component with the largest area was chosen as the best representative of abdominal contour. Since the acquired component was not necessarily a closed curve, the smallest convex–hull containing the entire component was calculated using a combination of the QuickHull and general dimension Beneath–Beyond algorithms *Barber et al.* (1996). The resulting convex hull was a coarse estimation of the shape and size of the abdomen, and its mass center was a coarse estimation of the abdominal location in an axial slice.

5.3.2.2 Orientation

As mentioned earlier, since there was a variation among different CT scans in terms of amount of tilt angle, it was necessary to find this orientation angle.

As the acquired convex hull was a good estimation of the abdominal region, it was used to find the orientation. After calculating the convex-hull, an ellipse was fitted to the result using the second moment of the area. The second moment of area for a simple polygon on a two dimensional plane can be computed by summing contributions from each edge of the polygon as:

$$\begin{aligned} I_x &= \frac{1}{12}\sum_{p=1}^{n}(y_p^2 + y_p y_{p+1} + y_{p+1}^2)(x_p y_{p+1} - x_{p+1} y_p) \\ I_y &= \frac{1}{12}\sum_{p=1}^{n}(x_p^2 + x_p x_{p+1} + x_{p+1}^2)(x_p y_{p+1} - x_{p+1} y_p) \\ I_{xy} &= \frac{1}{24}\sum_{p=1}^{n}(x_p y_{p+1} + 2x_p y_p + 2x_{p+1} y_{p+1} + x_{p+1} y_p) \\ &\quad \times (x_p y_{p+1} - x_{p+1} y_p) \end{aligned} \tag{5.1}$$

where i goes over all vertices, and $(x_p \ , \ y_p)$ (with $x_{n+1} = x_1$, $y_{n+1} = y_1$) are the coordinates of the polygon formed by connecting vertices of the convex hull.

By decomposing the matrix of second moments of area into the product of a rotation matrix Q and a diagonal matrix Λ, the properties of ellipse can be found.

$$I = \begin{bmatrix} I_x & I_{xy} \\ I_{xy} & I_y \end{bmatrix} = Q\Lambda Q^T. \tag{5.2}$$

5.3.2.3 Translation

The mass center of the acquired convex hull is a coarse estimation of the abdominal region center before the translation:

$$(X_{\text{center}-\text{old}}, Y_{\text{center}-\text{old}}) = (i_{cm}, j_{cm}) \tag{5.3}$$

where i_{cm} and j_{cm} are the mean of x-coordinate and y-coordinate respectively.

By placing the old center at the middle of the image, all images are aligned at the center. Assuming n_c and n_r respectively show the number of columns and rows, and are constant in all images, new center is calculated using:

$$(X_{\text{center}-\text{new}}, Y_{\text{center}-\text{new}}) = (\frac{\mathrm{n_c}}{2}, \frac{\mathrm{n_r}}{2}) \tag{5.4}$$

Having the old and new centers, each pixel needs to be shifted by $(\Delta x, \Delta y)$ where:

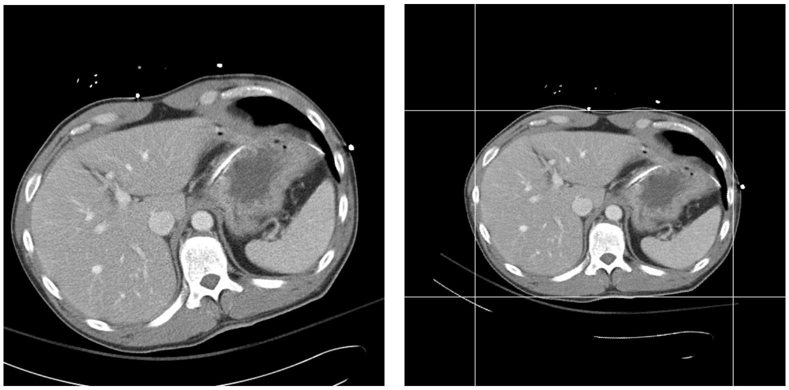
$$\begin{aligned} \Delta x &= X_{\text{center}-\text{new}} - X_{\text{center}-\text{old}} \\ \Delta y &= Y_{\text{center}-\text{new}} - Y_{\text{center}-\text{old}} \end{aligned} \tag{5.5}$$

5.3.2.4 Scaling

To detect the scale parameter in X and Y directions, width and height of the convex hull were calculated as W_{old} and H_{old}. By having the width and height of the reference shape as new scale, the scaling operation is defined as:

$$(x, y) \rightarrow (ax, ay) \tag{5.6}$$

where $a = (W_{\text{new}} + H_{\text{new}})/(W_{\text{old}} + H_{\text{old}})$. W_{new} and H_{new} correspond to the width and height of abdominal cavity in the reference images, respectively. Fig. 5.2 shows the original image before and after registration.



(a) (b)

Figure 5.2: Abdominal cavity alignment. (a) shows the original CT image before registration, and (b) shows the same image after registration (*Farzaneh et al.*, 2016).

5.3.3 Initialization Mask Detection

For segmentation purpose, we used ACM (Active Contour Modeling) whish is an efficient segmentation method. ACM starts with an initialization mask and evolves this mask throughout a number of iterations. In order to accurately segment the kidney using this technique, it is crucial to choose an initial mask accurately. To accomplish this, a machine learning algorithm was used to determine the probability of each pixel belonging to the kidney region. An adaptive thresholding method was then applied on the probability values to determine a 3D volume as the initial mask.

5.3.3.1 Machine Learning Classification Model

Next, potential discriminative features were extracted for each pixel in the abdominal cavity. However, extracting features for every single pixel is computationally intensive while providing redundant information. Thus, we first grouped sets of adjacent pixels in axial CT scans to build superpixels using the SLIC (Simple Linear Iterative Clustering) algorithm (*Achanta et al.*, 2012). Next, we selected a window W of 25×25 pixels, around the center of the mass of each superpixel. A set of location, histogram, and filtering-based features were extracted from each superpixel using its corresponding W (Table C.1). Histogram information included minimum, maximum, average, standard deviation, skewness, kurtosis, smoothness, and entropy of intensities of pixels within W. Textural information consisted of Gaussian, Laplacian of Gaussian, and Gabor filter-based features. Gabor features were calculated in eight evenly spaced orientations and four different frequencies. Laplacian of Gaussian highlights sharp intensity changes and is useful for edge detection. Finally, spatial features were determined as the location of a superpixel in the Cartesian coordinate system.

To determine the probability score of each superpixel belonging to the kidney region, a random forest classifier was trained using patient-wise 10 fold cross-validation.

5.3.3.2 Adaptive Probability Thresholding

Conventionally, in machine learning classification models a fixed threshold is applied as a cut-off on the probability score of each sample point to determine its class. However, due to heterogeneous representations of kidneys in CT scans, mainly due to various contrast phase and inter-patient anatomical variability, we did not apply a fixed threshold to the probability score. Instead, we applied an adaptive threshold model to segment out a 3D initialization volume inside the kidney. To calculate this threshold, T, we started with the maximum threshold value of 1 and gradually decrease the threshold value by Δ until we segmented out a 3D connected component larger than a predefined volume V_{min}. Therefore, the optimum cutoff threshold was computed by:

$$T = max\left\{1 - k\Delta | V_{1-k\Delta} > V_{min}, k = 0, 1, ..., \left\lfloor \frac{1}{\Delta} \right\rfloor \right\} \tag{5.7}$$

where V_T is the volume of the current largest connected component by applying cutoff value of T.

5.3.4 Kidney Surface Modeling using Active Contour Model

At this stage, the boundary of the initialization mask calculated in the previous section was evolved by using 3D active contour modeling. Active contour modeling evolved the initialization mask in an iterative process to be entirely constrained by the border of the desired object. As in many CT scans the edges between kidneys and neighboring organs such as liver were smooth, or injuries might have blurred the boundaries, edge-based active contour models were not effective. Thus, we implemented an active contour model known as Chan-Vese that was proposed to segment objects without well-defined edges (*Chan and Vese*, 2001). The Chan-Vese model is based on a level-set formulation and Mumford-Shah segmentation techniques, and is widely used in medical image processing.

In this work, after each fixed number of iterations (10), the evolved region was first refined based on intensity and then by its 3D representation (Fig. 5.3).

To refine the region based on intensity, the mean, μ, and standard deviation, σ, of the corresponding evolved region on the original CT image were calculated first. Then, pixels with intensity values outside $[\mu - \sigma, \mu + \sigma]$ were excluded from the segmentation. Due to the contrast phase or imaging settings, the pixels' intensity of kidney and liver might be very close to each other and hence part of the liver might

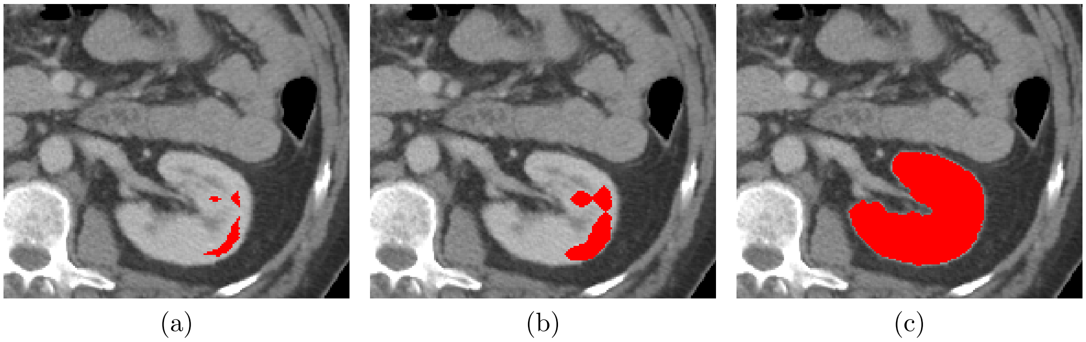


Figure 5.3: Evolving the kidney initialization mask iteratively. (a) The initialization mask is superimposed on the original image. (b) The result of mask evolution after 4 iterations. (c) Final segmentation (*Farzaneh et al.*, 2018).

be falsely segmented as kidney. Thus, approximate mean intensity values of kidneys and liver were used to find CT scans with such characteristics. Mean intensity of kidneys was calculated by averaging the pixel intensity of the current mask. In order to estimate the mean intensity of the liver, we used the liver probability atlas (see Appendix C for details) (*Farzaneh et al.*, 2016). We selected all pixels belonging to the liver with the highest probability and calculated their mean intensity value.

For a CT scan with low intensity contrast between liver and kidneys, a stricter condition was applied to exclude all pixels with intensity less than μ - 0.5 $\times$ σ (Fig. 5.4). Next, only pixels belonging to the largest connected component were selected as the kidney has a contiguous representation. Moreover, to prevent over-segmentation and minimize the running time, the volume of the current mask was calculated. If the volume was larger than a hard threshold, the iterative process was terminated. This threshold was based on the maximum of kidney volumes.

5.4 Results

We compared our segmentation results with the manually annotated ground truth verified by a radiologist. We used 10 fold cross-validation to train the random forest initialization mask detection model. Each fold included CT scans from 3 or 4

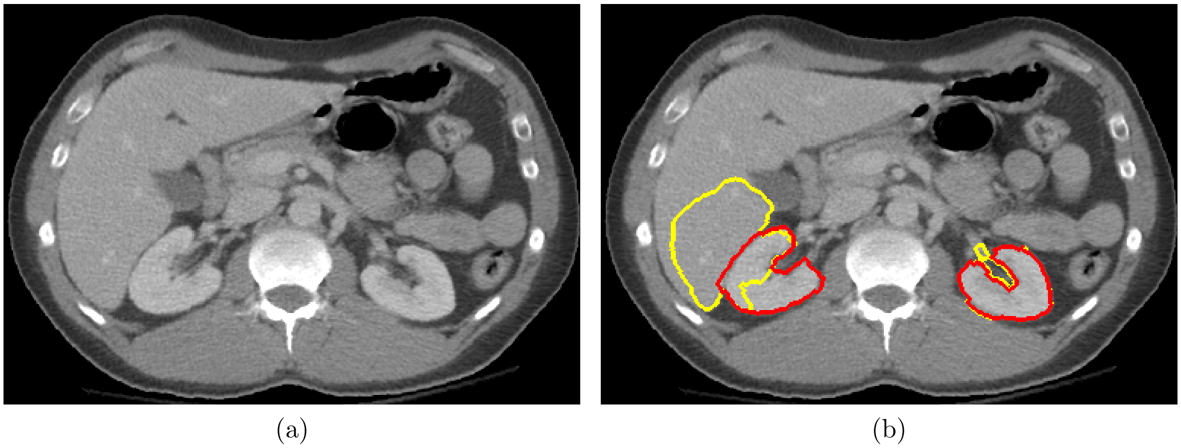


Figure 5.4: The effect of considering customized contrast information on kidney segmentation.(a) The original CT image exhibits low contrast between the kidneys and the liver. (b) The yellow contours show the result without considering the relative intensity distribution of liver and kidney in which the segmentation evolves into the liver. The red contours are the result after incorporating liver probability atlas information. The red contours overlay the yellow ones (*Farzaneh et al.*, 2018).

patients. Except for the left kidney in one patient who was in shock, our algorithm correctly detected the initialization mask entirely within the kidney region. Our final segmentation results show pixel-wise Dice, recall, precision value of 88.9% , 92.6% , and 86.4%, respectively.

Table 5.1: Performance of our proposed kidney segmentation approach stratified based on the kidney location. Numbers in parentheses are standard deviations.

Segmentation Result	**Dice (%)**	**Recall (%)**	**Precision (%)**	**RVD (%)**	**VOE (%)**
Left kidney	88.58 (16.55)	90.81 (17.71)	87.20 (17.05)	4.19 (16.21)	18.09 (17.11)
Right kidney	89.03 (8.12)	94.50 (7.03)	85.56 (12.49)	13.87 (25.52)	18.91 (12.14)

Fig. 5.5 shows variations caused by contrast phase and trauma and the resultant segmentations.

Table 5.2 compares the performance of our model and state-of-the-art methods. Note that each model is developed on a different data set with different imaging settings and potentially different populations of interest, thus direct comparison cannot

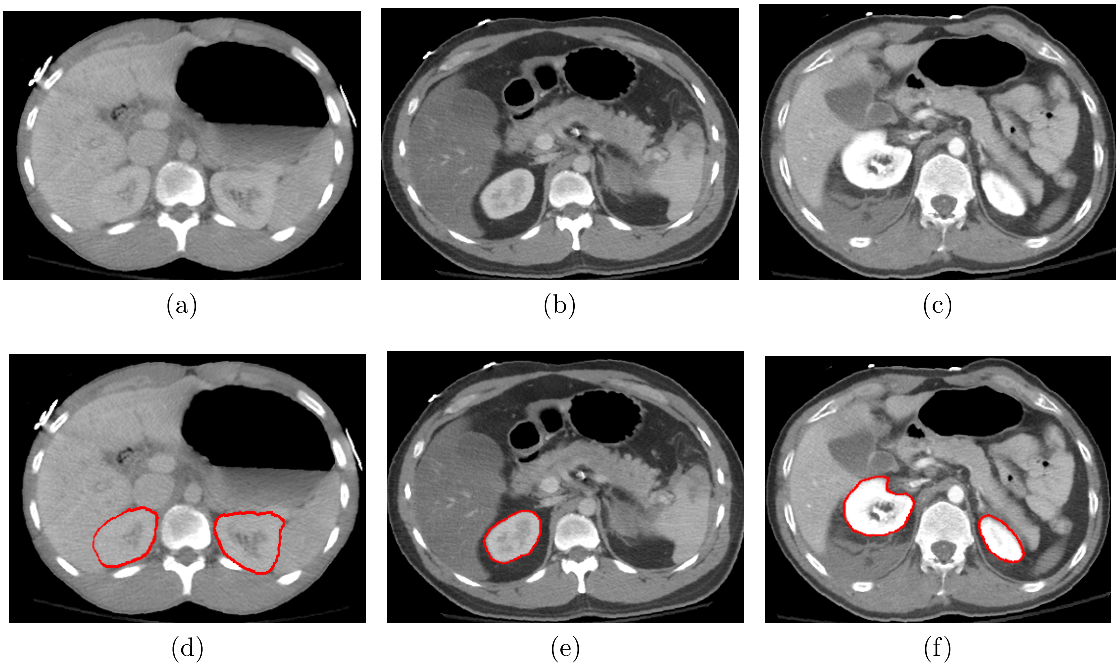


Figure 5.5: Kidney segmentation result for CT images with various contrasts between the liver and kidneys. (a)-(c) CT images with various contrasts between adjacent kidney and liver. (d)-(e) The red marks represent the ground truth kidneys contours (*Farzaneh et al.*, 2018).

be made from this table. For example, Khalifa et al. developed their method on CT images collected from 20 subjects while each subject has high resolution CT scan (slice thickness of 0.9 *mm*) at 3 known contrast phases as pre-, post-, and delayed contrast phases (*Khalifa et al.*, 2017).

Moreover, although our results are promising, in one severe case our algorithm was unable to segment cysts - non-trauma pre-existing abnormal fluid-filled sacs; while in ground truth these were considered as parts of the kidneys (Fig. 5.6).

5.5 Conclusion

In this chapter, we investigated the use of machine learning and active contour modeling in kidney segmentation. A random forest classifier was employed to detect a high-probability initialization mask inside the kidney. This initialization was then

Table 5.2: Performance of our proposed kidney segmentation approach compared with other methods in the literature. Numbers in parentheses are standard deviations. For the cited studies, scores are reported as presented in the original papers.

Method	**Patient Population**	**Contrast Phase**	**CT Slice Spacing (Resolution along the Z-axis)**	**Dice (%)**
Proposed Method (n=35)	Blunt abdominal trauma	Different and unknown	5 mm	88.88 (9.34)
(*Lin et al.*, 2006) (n=30)	Pathological	Not reported	8-10 mm	88 (2.57)
(*Khalifa et al.*, 2017) (n=20)	Not reported	Three known contrast phases for each patient	0.9 mm	97.27 (0.83)
(*Skalski et al.*, 2017) (n=10)	Renal cancer	Known	1-3 mm	86.2 (3.9)

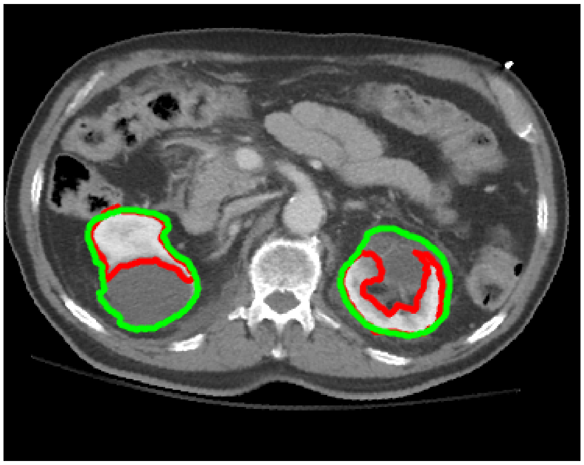


Figure 5.6: Kidney segmentation result for a patient with kidney cyst. The green contours represent the ground truth (including cysts). The red contours are the algorithm final result. The ground truth overlays some parts of segmented result (*Farzaneh et al.*, 2018).

iteratively evolved using active contour modeling to fit the boundary of the kidney. The process was performed independently for each kidney. Since the right kidney is adjacent to the liver, the resulting segmentation was refined after each batch of iterations with respect to the intensity distribution of the segmented right kidney and the liver.

Unlike previous methods, our algorithm was developed and evaluated on CT images collected from trauma patients at dissimilar contrast phases. Except for the left kidney in one patient who was in shock, the initialization mask was placed entirely inside the kidney parenchyma for all instances. Considering a high degree of variation (biological or induced by contrast phase, pre-existing conditions, and trauma injuries), the result shows robustness and generalizability of the proposed machine learning classifier in localizing kidneys.

In the future, when a larger data set is available, the applicability of deep convolutional neural network models can be investigated. Moreover, with access to enough kidney trauma cases, this type of injury can be automatically detected and assessed in future works.

CHAPTER VI

Concluding Remarks and Future Directions

Trauma is a major worldwide public health problem and the leading cause of death among Americans younger than 46. Traumatic brain injuries and abdominal trauma are among the major contributors to traumatic mortality and morbidity. Early diagnosis of traumatic injuries and development of an accurate prognosis are critical for optimal triage and patient management. Current trauma protocols utilize CT assessment of injuries in a subjective and qualitative (vs. quantitative) fashion, shortcomings that could both be addressed by automated computer-aided systems capable of generating real-time, reproducible and quantitative information.

This study outlines multiple frameworks for quantitative diagnosis and prognosis of traumatic injuries. The overarching focus of this study was to infuse clinical domain knowledge in the state-of-the-art artificial intelligence black box models to achieve greater generalizability and trustworthiness.

In this chapter, we summarize the key achievements of this dissertation and discuss potential directions for future research.

6.1 Predicting the Long-Term Functional Outcome of Traumatic Brain Injury Patients

In Chapter II, a machine learning framework was proposed that enabled the creation of a robust explainable model for individualized prediction of TBI functional outcomes as defined by GOSE. The proposed framework is transparent with respect to understanding how input variables result in the model's decision at both the individual and patient population levels. To achieve the robustness, a bootstrapping method was employed to only select the variable with consistent population-level contribution behavior. Such a model can achieve high accuracy, avoid collinearity-induced biases, and ultimately accelerate adoption of machine learning models in clinical settings.

6.1.1 Main Findings

Among 62 candidate variables from EHR, 18 were identified to be robust and have a contribution behavior aligned with clinical domain knowledge for TBI prognostication. Among 18 variables included in the final model, both age and the number of brain regions with subarachnoid hemorrhage were ranked as the most impactful variables in predicting TBI outcome. These variables were followed by GCS motor response, intra-ventricular hemorrhage, GCS eye opening, third ventricle compression, and SDH. The four laboratory values determined to be appropriate predictors were hemoglobin, glucose, activated Partial Thromboplastin Time (aPTT), and International Normalized Ratio (INR). Overall, the results showed that imposing the statistical and expert "checks and balances" did not adversely affect model performance.

6.1.2 Future Directions

In the future, this framework can be applied to similar tasks to determine potential sources of bias in a black box model as indicated by counterintuitive population-level contribution behaviors. Understanding biases in a predictive model is of a crucial importance, specifically if they are driven by self-reinforcing positive feedback. For example, patients with a particular pre-existing condition might receive a more aggressive level of care that can potentially lead to a favorable outcome. As a result, a machine learning model might learn the pattern that the existence of the pre-existing condition, regardless of treatment plan, lowers the risk of unfavorable outcome. Once applied in a real-world setting, this self-reinforcing positive feedback can result in patients with the particular pre-existing condition receiving less aggressive care than they would have otherwise.

6.2 Automated Detection and Severity Assessment of Subdural Hematoma in Traumatic Brain Injury Patients

In Chapter III, we proposed a fully automated approach for segmentation and radiographic severity assessment of SDH by analyzing head CT scans. First, a SDH segmentation model was developed by integrating hand-crafted features with data-driven deep features. Based on the automatically measured hematoma volume, we categorized SDH patients as severe and non-severe. This model enables the accurate quantification of blood, which is otherwise almost impossible due to the time-consuming manual process.

6.2.1 Main Findings

Compared to deep learning models, integrating hand-crafted features with data-driven deep features leads to a higher overall accuracy as well as greater consistency and robustness in segmenting SDH regions. Moreover, the proposed model was shown to be more generalizable to different types and severity levels of SDHs. The primary reason for this result is that the hand-crafted features reflect human domain knowledge that can compensate for limitations in deep model performance on unobserved regions of the input data distribution.

6.2.2 Future Directions

The proposed model was trained on CT scans captured using LightSpeed VCT or Discovery CT750 HD systems, both from GE, the U-M vendor of choice. CT scans from other manufacturers could be incorporated into future models to create a vendor-agnostic model.

We should recognize that the proposed algorithm should be tested and validated against a larger population from an external data set and obtain regulatory approval and clinician acceptance before it can be incorporated into clinical practice. The predictive power of the SDH volumetric measurement should also be evaluated against a representative data set that covers different variations of SDH in traumatic brain injury patients.

6.3 Automated Detection and Severity Assessment of Liver Trauma

This study is the first to automatically identify and assess liver parenchymal traumas utilizing contrast-enhanced CT without taking advantage of any prior knowledge about the presence of the injury. We developed a fully automated framework capable of providing objective and quantitative information about the presence and extent of liver trauma. For this task, deep convolutional neural networks are employed to generate liver parenchyma and trauma segmentation masks. Next, human domain knowledge regarding the location and intensity distribution of liver trauma was integrated into the model to refine the initial segmentation masks. The severity of trauma was quantified as the proportion of the liver tissue affected by those injuries (i.e., liver disruption involvement).

6.3.1 Main Findings

Non-traumatic pre-existing conditions, such as fatty liver and congestive hepatopathy, may affect the liver parenchyma's appearance in CT scans. As a result, the injury assessment for patients with these pre-existing conditions (minority class) might suffer from automation bias due to less representation in the training phase. In this study, we showed that human domain knowledge regarding the location and intensity distribution of liver trauma can effectively address such bias to achieve a more generalizable model.

The high performance of the model in both assessing blunt liver trauma and avoiding false positives in non-trauma (control) patients supports its potential applicability to enhance the clinical decision-making process.

6.3.2 Future Directions

One of the study's limitations is its low sample size, which leads to a lack of enough massive traumas or CT scans with strong beam hardening artifacts in the training phase. Due to low representation, the model is not effectively optimized for such cases. Further improvements can be achieved through more extensive studies in which massive trauma and non-trivial image artifacts are adequately represented.

In future studies, other major types of liver trauma, including subcapsular hematoma and active arterial bleeding, should also be automatically assessed to develop a comprehensive liver trauma assessment tool. Automated detection of active arterial bleeding would be a critical

6.4 Automated Kidney Segmentation for Patients with Abdominal Trauma

Chapter V focuses on segmenting kidneys for patients with blunt abdominal trauma. We developed a framework using machine learning classifiers that identify pixels belonging to kidneys with high probabilities. These initial masks are then evolved to fit the boundary of the kidney without crossing the border between kidney and liver. Localizing kidneys is a challenging task due to the heterogeneous representation of kidneys on CT scans. This variation can be attributed to factors such as normal biological variations (e.g., displacement of kidneys), contrast phase, pre-existing conditions (e.g., cyst), or potential injuries (e.g., laceration, shock). Considering a high degree of variations, the result shows robustness and generalizability of machine learning classifiers in localizing kidneys.

6.4.1 Future Directions

The applicability of deep learning models can be investigated by expanding the data set in the future.

The small number of cases with kidney trauma prevented us from developing models for detection and severity assessment of such injuries. Thus, the analysis was limited to kidney segmentation. As more kidney trauma data is collected, future investigations could focus on segmentation and quantification of this type of trauma as well.

6.5 Perspective on Image Processing Model Selection

Different choices of image processing models were employed in this dissertation research depending on the underlying assumptions of each problem. The method used for subdural hematoma segmentation that incorporates a joint feature representation of hand-crafted and data-driven is shown to generate more accurate and robust initial segmentation masks compared to U-net alone. However, this model was not applicable in analyzing abdominal CT scans due to the fact that extracting hand-crafted features for large regions of interest, such as the abdominal cavity (vs. intracranial region), is time-consuming, leading to models that might be incapable of real-time analysis. Thus, U-net alone was applied to generate the initial masks for the liver parenchymal and trauma segmentation task. Thus, in time sensitive applications, for small volumes, we propose to use the combinatory approach proposed in Chapter III, and for larger volumes the deep convolutional neural network architectures alone are preferable.

A critical step in developing reliable image processing techniques for health-care applications is an extensive investigation of the potential effects of pre-existing conditions on radiographic images and the final performance. This a critical process to systematically post-process segmentation maps.

6.6 Future Perspective on Polytrauma Assessment

In the future, the generalizability of each of the proposed models needs to be validated on an external data set. This was not undertaken in this study due to limited data availability. However, we believe that incorporating human domain knowledge into the model can potentially compensate for the lack of generalizability inherent in most black box machine learning and deep learning methods. Involving clinicians in the model development process can help eliminate biases and error modes from a model, thereby increasing the likelihood that physicians will utilize the model. Moreover, the utility of the proposed models needs to be determined prospectively in a clinical setting prior to its incorporation into the standard practice. These systems' utilities can potentially be tested with clinicians of different ages and experience levels since these characteristics might indicate how comfortable a clinician is with using a new automated method and rate how useful it was beyond one's usual practice.

Overall, this thesis is part of a greater effort for a comprehensive polytrauma decision support system. In addition to liver, kidney, and brain subdural hematoma, the polytrauma decision support system comprises assessment of acute intracranial hemorrhage, spleen trauma, and pelvic bone fracture. Together, these components contribute to improving trauma diagnosis and prognosis accuracy, and help prevent delayed- or missed-diagnoses. Our long-term goal is to develop a fully automated system that can provide quantitative trauma severity assessments and a personalized treatment plan.

APPENDICES

APPENDIX A

Additional Information for Chapter II

Machine Learning Algorithm Selection

To select the machine learning algorithm to be used in this study, we trained five different algorithms including XGBoost, deep learning, logistic regression, and support vector machine using the initial 62 candidate features. For deep learning, we used feed forward neural network with 4 hidden layers. The model that performed the best on the validation set is then chosen for the remainder of the process. Table A.2 compares the performance of the candidate machine learning algorithms. Although the performance on the validation is used to select the model, for more information, we included the performance on the training and test sets in Table A.2 as well. XGBoost model outperformed the other methods and was chosen in our TBI prognostication study.

Table A.1: List of candidate and selected variables for traumatic brain injury outcome prediction and their definitions. [a] Subarachnoid hemorrhage refers to bleeding into the subarachnoid space between the brain and the surrounding membrane. Brain regions include suprasellar, basal cisterns, right and left Sylvian fissure, right and left interhemispheric, right and left lobar-frontal, right and left lobar-parietal, right and left lobar-occipital, right and left lobar temporal. [b] Intraparenchymal hemorrhage refers to bleeding within the brain parenchyma. Brain regions include midbrain/pons, right and left frontal, right and left temporal, right and left parietal, right and left occipital, right and left basal ganglia, right and left posterior fossa. [c] Brain contusions refer to the bruises of the brain tissue. Brain regions include midbrain/pons, right and left frontal, right and left temporal, right and left parietal, right and left occipital, right and left basal ganglia, right and left posterior fossa. [d] Diffuse Axonal Injury (DAI) corresponds to shearing of the brain's axons due to brain shifts or rotations after an injury. Brain regions include right and left frontal, right and left parietal, right and left basal ganglia, brainstem, corpus callosum, right and left centrum semiovale. [e] Brain regions include midbrain/pons, right and left frontal, right and left temporal, right and left parietal, right and left occipital, right and left basal ganglia.

Name	**Definition (unit)**	**Median** **(min–max)**	**No: Yes or** **None: One: Two**	**Selected in** **Final Model**
Demographics				
Age		35 (17-94)		Yes
Sex: female			607 (73.04%): 224 (26.96%)	
Baseline features				
Best motor response	As defined by (*Teasdale and Jennett*, 1976)	4 (1-6)		Yes

Table A.1 continued from previous page

Name	Definition (unit)	Median (min–max)	No: Yes or None: One: Two	Selected in Final Model
Best eye opening response	As defined by (*Teasdale and Jennett*, 1976)	1 (1-4)		Yes
Best verbal response	As defined by (*Teasdale and Jennett*, 1976)	1 (1-5)		Yes

Pupil response	None, one, or both eyes		35 (4.2%):
			125 (15.0%):
			671 (80.7%)	

Radiology report				

Epidural hematoma (#)	Zero if none, one if unilateral, and two if bilateral epidural hematoma		710 (85.44%):
			110(13.24%):
			11(1.32%)	

Epidural hematoma (max width)	(mm)	0 (0-102)		

Table A.1 continued from previous page

Name	**Definition (unit)**	**Median (min–max)**	**No: Yes or None: One: Two**	**Selected in Final Model**

Subdural hematoma (#)	Zero if none, one if unilateral, and two if bilateral subdural hematoma		425 (51.1%):
			330 (39.7%):
			76 (9.1%)
				Yes

Subdural hematoma (max width)	(mm)	0 (0-125)		Yes
Subarachnoid hemorrhage (#)	Number of brain regions with subarachnoid hemorrhage[a]	1 (0-14)		Yes

Intra-ventricular hemorrhage	Zero if none, one if minimal layering, and two if clot intra-ventricular hemorrhage		642 (77.3%):
			119 (14.3%):
			70 (8.4%)
				Yes

Intraparenchymal hematoma (#)	Number of brain regions with intraparenchymal hemorrhage[b]	0 (0–4)		
Intraparenchymal hematoma (max width)	(mm)	0 (0-67)		Yes

Table A.1 continued from previous page

Name	**Definition (unit)**	**Median (min–max)**	**No: Yes or None: One: Two**	**Selected in Final Model**
Evidence of surgical evacuation	Evidence of surgical evacuation of intraparenchymal hematoma		827 (99.52%): 4 (%0.48)	
Brain contusion (#)	Number of brain regions with brain contusion [c]	0 (0-5)		Yes
Brain contusion (max width)	(mm)	0 (0-93)		Yes
DAI finding (#)	Number of brain regions with diffuse axonal injury [d]	0 (0-6)		Yes
DAI finding (max width)	Maximum width of diffuse axonal injury (mm)	0 (0–22)		

Generalized edema severity	Zero if none, one of mild, and two if moderate edema		654 (78.70%):
			52 (6.26%):
			125 (15.04%)	

Table A.1 continued from previous page

Name	**Definition (unit)**	**Median (min–max)**	**No: Yes or None: One: Two**	**Selected in Final Model**
Focal swelling (#)	Number of brain regions with focal swelling ^{e}	0 (0-3)		
Midline shift	Shift of over 5 mm		705 (84.84%): 126 (15.16%)	

Sulcal obliteration	Zero if none, one if unilateral, and two if bilateral sulcal obliteration		643 (77.38%):
			77 (9.27%):
			111 (13.36%)	

Lateral ventricle compression			835 (76.41%): 159 (19.13%)	
Third ventricle compression			653 (78.6%):: 178 (21.4%)	Yes
Transtentorial herniation			700 (84.2%): 131 (15.8%)	Yes

Table A.1 continued from previous page

Name	Definition (unit)	Median (min–max)	No: Yes or None: One: Two	Selected in Final Model
Uncal herniation			714 (85.92%): 117 (14.08%)	
Tonsillar herniation			767 (92.30%): 64 (7.70%)	
Upward herniation			819 (98.56%): 12 (1.44%)	
Depressed skull fracture			773 (93.02%): 58 (6.98%)	
Basilar skull fracture			646 (77.74%): 185 (22.26%)	
Abbreviated Injury Scores				
Neck		4 (0-6)		

Table A.1 continued from previous page

Name	**Definition (unit)**	**Median (min–max)**	**No: Yes or None: One: Two**	**Selected in Final Model**
Face		0 (0–4)		
Chest		1 (0-5)		
Abdomen		0 (0–5)		
Extremity		1 (0–5)		
External skin		1 (0–4)		
Laboratory values				
Glucose	(mg/dL)	143 (68-554)		Yes

Table A.1 continued from previous page

Name	**Definition (unit)**	**Median** **(min–max)**	**No: Yes or** **None: One: Two**	**Selected in** **Final Model**
Creatinine	(mg/dL)	1.0 (0.3-4.2)		
Potassium	(mmol/L)	3.7 (1.5-6.5)		
Sodium	(mmol/L)	140 (125-157)		
Chloride	(mmol/L)	105 (88-130)		
Bicarbonate	(mmol/L)	23 (8-34)		
Hgb	Hemoglobin (g/dL)	13.8 (2.0-18.7)		Yes
WBC	White blood cell count (×10^9/L)	13.6 (3.2-41.4)		

Table A.1 continued from previous page

Name	**Definition (unit)**	**Median (min–max)**	**No: Yes or None: One: Two**	**Selected in Final Model**
Platelets	Platelet count ($\times 10^3/\mathrm{mm}^3$)	237 (36-700)		
aPTT	Activated partial thromboplastin time (sec)	26 (12-73)		Yes
INR	International Normalized Ratio	1.1 (0.8-12.0)		Yes
Medical history				
Active neurological disease	Includes prior TBI hospitalization or medical evaluation, CVA, seizure, paralysis/neurological weakness, headache, sleep disorder, and other unknown		804 (96.75%): 27 (3.25%)	

Table A.1 continued from previous page

Name	**Definition (unit)**	**Median (min–max)**	**No: Yes or None: One: Two**	**Selected in Final Model**
Inactive neurological disease	Prior neurological disease		735 (88.45%): 96 (11.55%)	
Active cardiovascular disease	Includes heart disease, hypertension, arrhythmias, and other unknown		692 (83.3%): 139 (16.7%)	
Inactive cardiovascular disease	Prior cardiovascular disease		787 (94.71%): 44 (5.29%)	
Active pulmonary disease	Includes COPD or asthma, and other unknown		779 (93.73%): 52 (6.26%)	
Inactive pulmonary disease	Prior pulmonary disease		793 (95.43%): 38 (0.4.57%)	
Active metabolic disease	Includes diabetes mellitus, pituitary disease, and other unknown		760 (91.46%): 71 (8.54%)	

Table A.1 continued from previous page

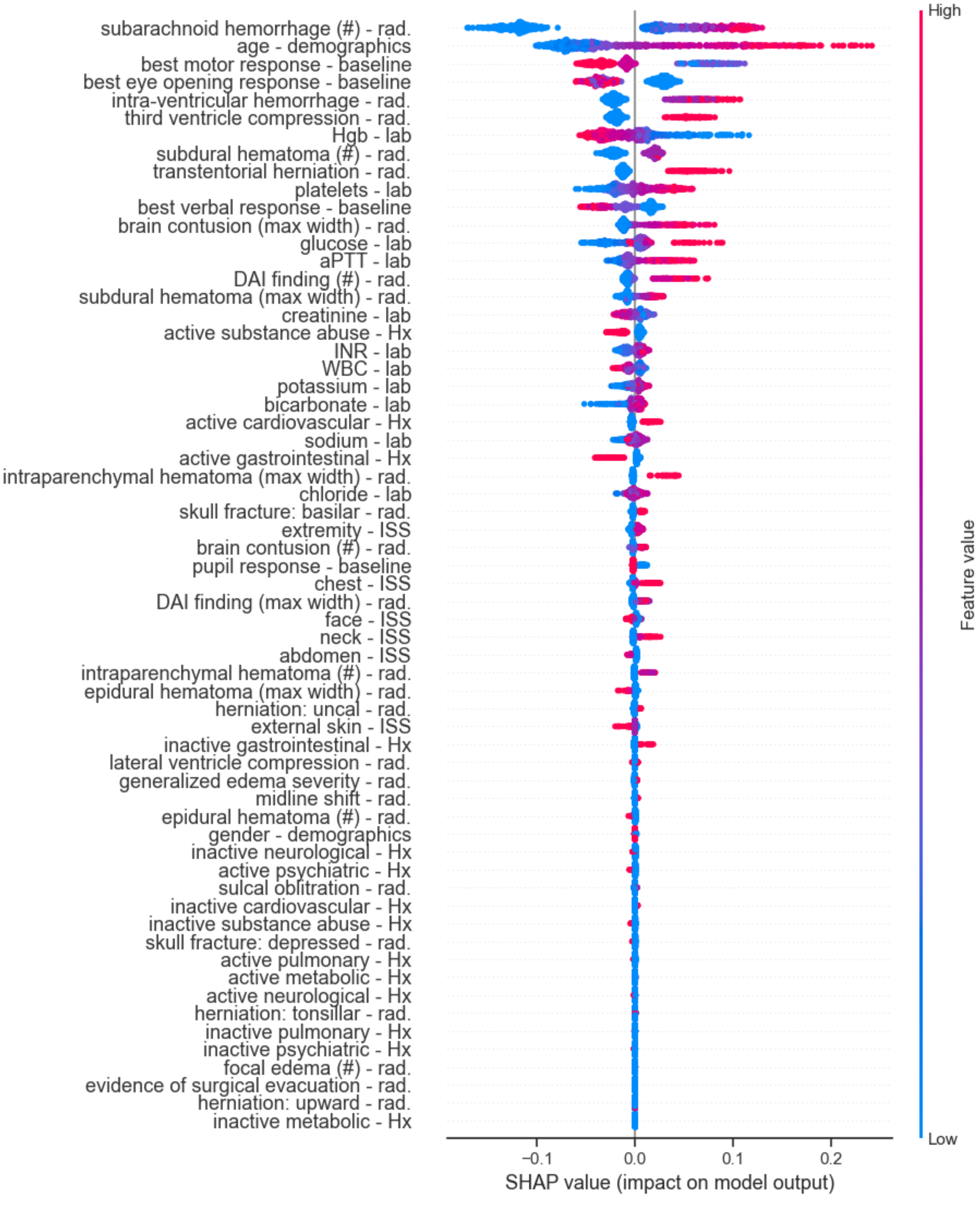
Name	**Definition (unit)**	**Median (min–max)**	**No: Yes or None: One: Two**	**Selected in Final Model**
Inactive metabolic disease	Prior metabolic disease		827 (99.52%): 4 (0.48%)	
Active gastrointestinal disease	Includes liver disease, hepatitis, and other unknown		768 (92.42%): 63 (7.58%)	
Inactive gastrointestinal disease	Prior gastrointestinal disease		800 (96.27%): 31 (3.73%)	
Active psychiatric disease	Includes depression/ suicidal gestures, schizophrenia, anxiety, and other unknown		699 (84.12%): 132 (15.88%)	
Inactive psychiatric disease	Prior psychiatric disease		802 (96.51%): 29 (3.49%)	
Active substance abuse	Alcohol and non-prescribed drug abuse		592 (71.24%): 239 (28.76%)	

Table A.1 continued from previous page

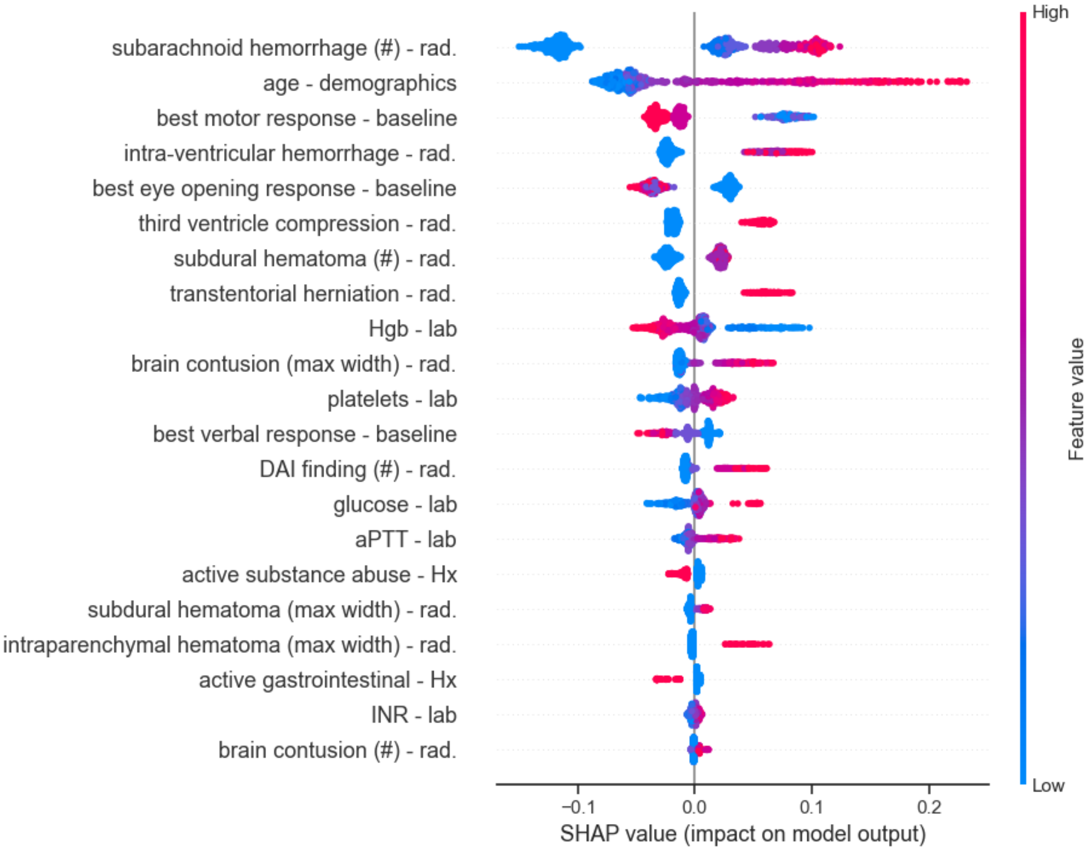
Name	**Definition (unit)**	**Median (min–max)**	**No: Yes or None: One: Two**	**Selected in Final Model**
Inactive substance abuse	Prior substance abuse		782 (94.10%) 49 (5.90%)	

Table A.2: Comparing the performance of multiple machine learning algorithms in predicting GOSE < 4 using initial candidate variables. Numbers in parentheses are standard deviations. Standard deviation is calculated over 5 cross-validation folds.

Training Set				
Method	**XGBoost**	**Deep learning**	**Logistic regression**	**Support vector machine**
AUC (%)	93.72 (2.36)	94.03 (4.82)	86.81 (0.43)	81.93 (1.35)
Accuracy (%)	85.22 (3.27)	87.02 (7.03)	78.68 (1.36)	74.68 (2.40)
F1 (%)	82.81 (3.60)	84.67 (7.47)	75.27 (1.06)	66.13 (4.09)
Sensitivity (%)	84.77 (3.05)	83.24 (4.41)	77.49 (1.87)	59.39 (5.45)
Specificity (%)	85.54 (4.40)	89.74 (10.12)	79.54 (3.19)	85.68 (1.16)
Precision (%)	81.06 (5.29)	86.77 (11.94)	73.29 (2.73)	74.79 (2.26)
Validation Set				
Method	**XGBoost**	**Deep learning**	**Logistic regression**	**Support vector machine**
AUC (%)	78.22 (1.26)	76.07 (2.69)	76.52 (3.04)	78.38 (2.06)
Accuracy (%)	75.00 (1.69)	70.51 (3.98)	71.64 (2.25)	75.00 (3.34)
F1 (%)	71.29 (1.90)	65.40 (3.64)	67.35 (3.77)	66.52 (4.92)
Sensitivity (%)	74.34 (4.89)	66.69 (7.62)	70.55 (8.28)	59.78 (6.95)
Specificity (%)	75.49 (4.56)	73.30 (9.63)	72.48 (5.42)	85.96 (4.79)
Precision (%)	68.80 (3.30)	65.19 (6.53)	65.05 (3.03)	75.95 (6.02)
Test Set				
Method	**XGBoost**	**Deep learning**	**Logistic regression**	**Support vector machine**
AUC (%)	80.94	77.90	80.33	76.95
Accuracy (%)	75.36	72.90	73.91	70.05
F1 (%)	70.52	65.00	67.47	58.11
Sensitivity (%)	70.11	59.77	64.37	49.43
Specificity (%)	79.17	82.50	80.83	85.00
Precision (%)	70.93	71.23	70.89	70.49

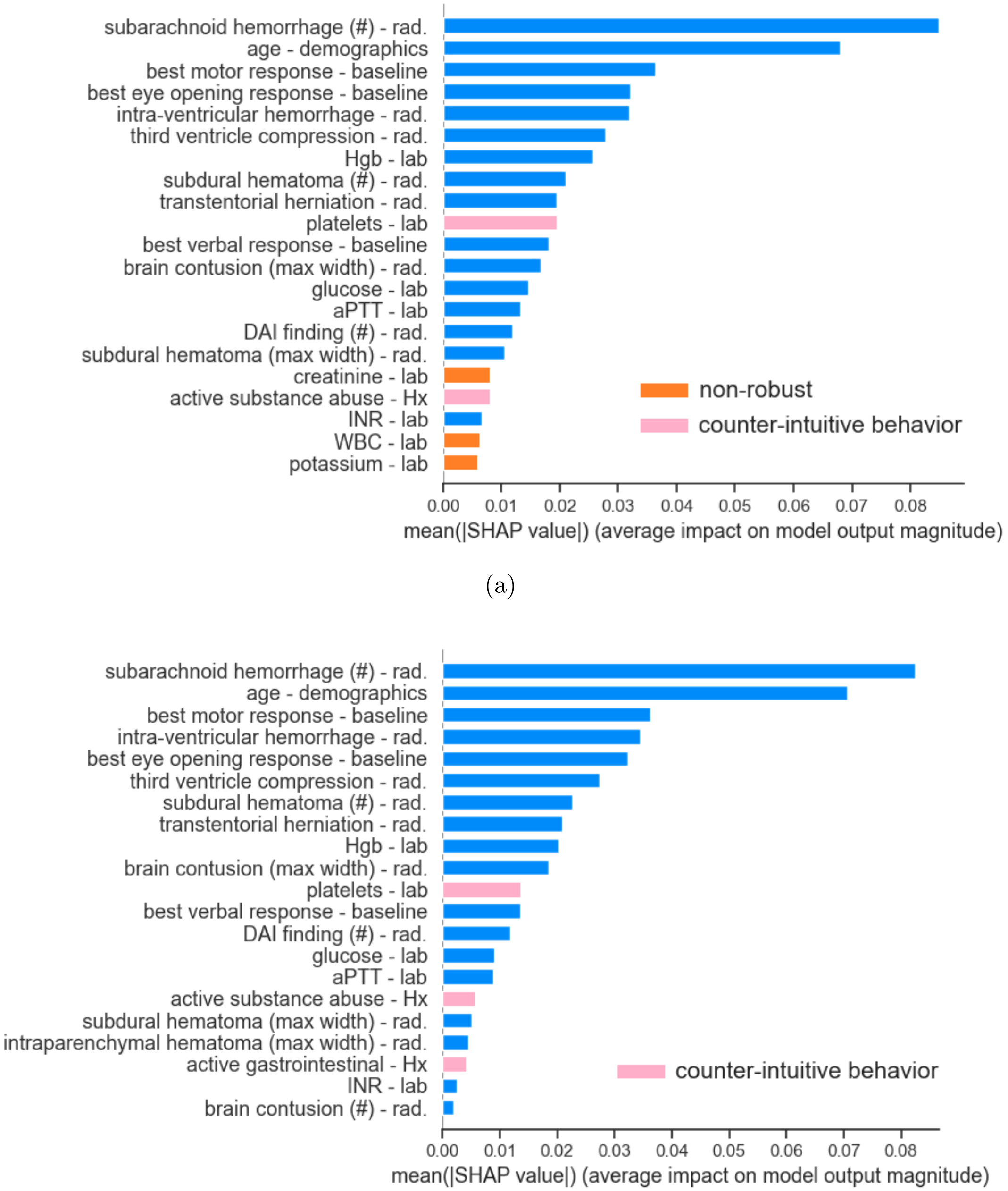

subarachnoid hemorrhage (#) - rad.
age - demographics
best motor response - baseline
best eye opening response - baseline
intra-ventricular hemorrhage - rad.
third ventricle compression - rad.
Hgb - lab
subdural hematoma (#) - rad.
transtentorial herniation - rad.
platelets - lab
best verbal response - baseline
brain contusion (max width) - rad.
glucose - lab
aPTT - lab
DAI finding (#) - rad.
subdural hematoma (max width) - rad.
creatinine - lab
active substance abuse - Hx
INR - lab
WBC - lab
potassium - lab
bicarbonate - lab
active cardiovascular - Hx
sodium - lab
active gastrointestinal - Hx
intraparenchymal hematoma (max width) - rad.
chloride - lab
skull fracture: basilar - rad.
extremity - ISS
brain contusion (#) - rad.
pupil response - baseline
chest - ISS
DAI finding (max width) - rad.
face - ISS
neck - ISS
abdomen - ISS
intraparenchymal hematoma (#) - rad.
epidural hematoma (max width) - rad.
herniation: uncal - rad.
external skin - ISS
inactive gastrointestinal - Hx
lateral ventricle compression - rad.
generalized edema severity - rad.
midline shift - rad.
epidural hematoma (#) - rad.
gender - demographics
inactive neurological - Hx
active psychiatric - Hx
sulcal oblitration - rad.
inactive cardiovascular - Hx
inactive substance abuse - Hx
skull fracture: depressed - rad.
active pulmonary - Hx
active metabolic - Hx
active neurological - Hx
herniation: tonsillar - rad.
inactive pulmonary - Hx
inactive psychiatric - Hx
focal edema (#) - rad.
evidence of surgical evacuation - rad.
herniation: upward - rad.
inactive metabolic - Hx
−0.1
0.0
0.1
0.2
SHAP value (impact on model output)
High
Feature value
Low

(a)



(a)

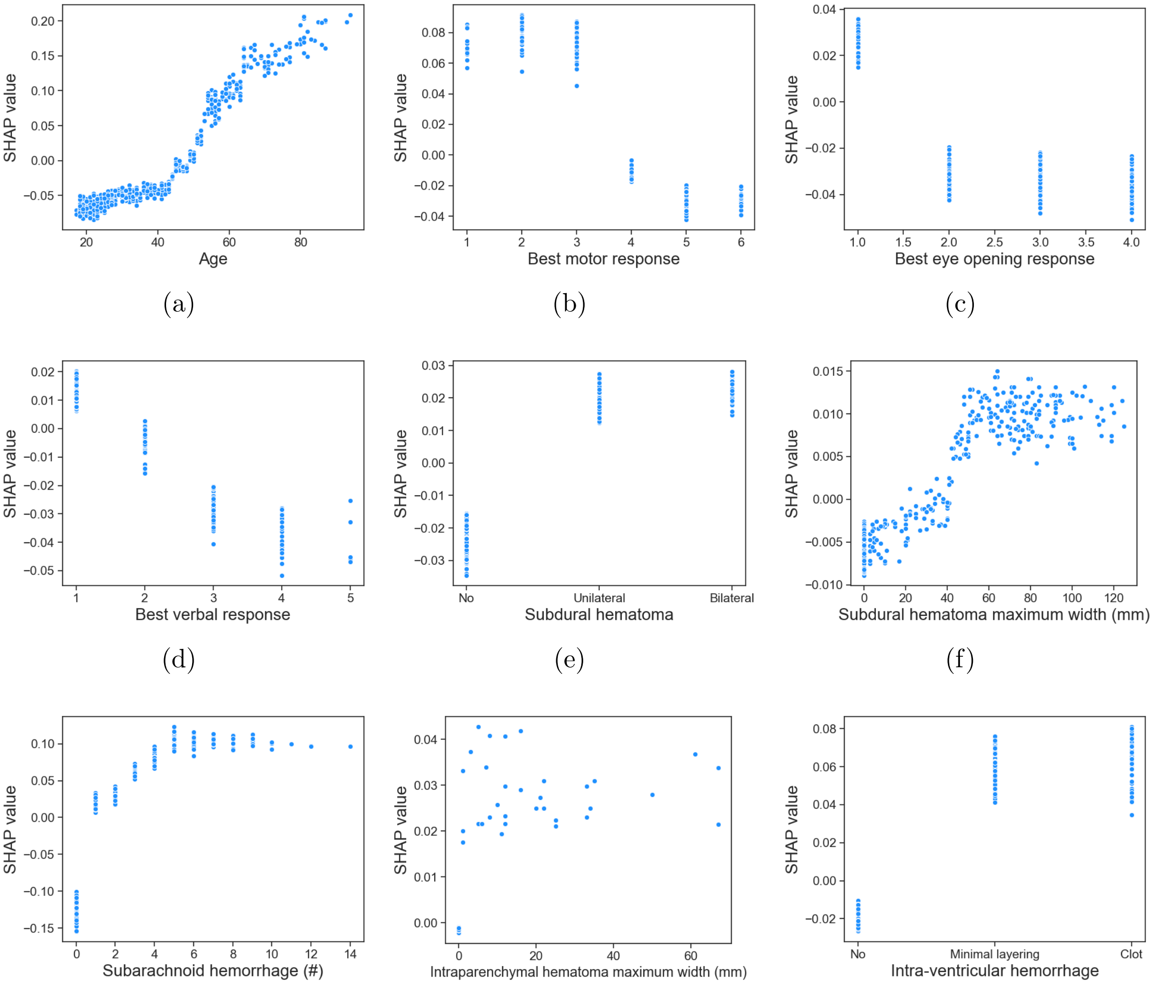
Figure A.1: Summary of SHAP contribution in the initial and intermediate model. (a) shows the summary of contributions in a model trained using the 62 candidate variables, (b) shows the summary of contributions in a model trained using the selected 21 robust variables. Variable types are denoted as follows - *rad*: radiology report, *lab*: laboratory value, *Hx*: medical history, and *ISS*: injury severity score (*Farzaneh et al.*, 2021b).



(a)

(b)

Figure A.2: SHAP importance plot. Variables are shown in order of their average impact on the predicted risk, where impact is defined as the average absolute SHAP value. (a) corresponds to the initial model with 62 candidate features. Only 21 variables with the highest impact are shown in the plot. (b) corresponds to the model with 21 robust variables. Variable types are denoted as follows - *rad*: radiology report, *lab*: laboratory value, *Hx*: medical history, and *ISS*: injury severity score (*Farzaneh et al.*, 2021b).


SHAP value
0.20
0.15
0.10
0.05
0.00
−0.05
20
40
60
80
Age
SHAP value
0.08
0.06
0.04
0.02
0.00
−0.02
−0.04
1
2
3
4
5
6
Best motor response
SHAP value
0.04
0.02
0.00
−0.02
−0.04
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Best eye opening response
SHAP value
0.02
0.01
0.00
−0.01
−0.02
−0.03
−0.04
−0.05
1
2
3
4
5
Best verbal response
SHAP value
0.03
0.02
0.01
0.00
−0.01
−0.02
−0.03
No
Unilateral
Bilateral
Subdural hematoma
SHAP value
0.015
0.010
0.005
0.000
−0.005
−0.010
0
20
40
60
80
100
120
Subdural hematoma maximum width (mm)
SHAP value
0.10
0.05
0.00
−0.05
−0.10
−0.15
0
2
4
6
8
10
12
14
Subarachnoid hemorrhage (#)
SHAP value
0.04
0.03
0.02
0.01
0.00
0
20
40
60
Intraparenchymal hematoma maximum width (mm)
SHAP value
0.08
0.06
0.04
0.02
0.00
−0.02
No
Minimal layering
Clot
Intra-ventricular hemorrhage

(a)

(b)

(c)

(d)

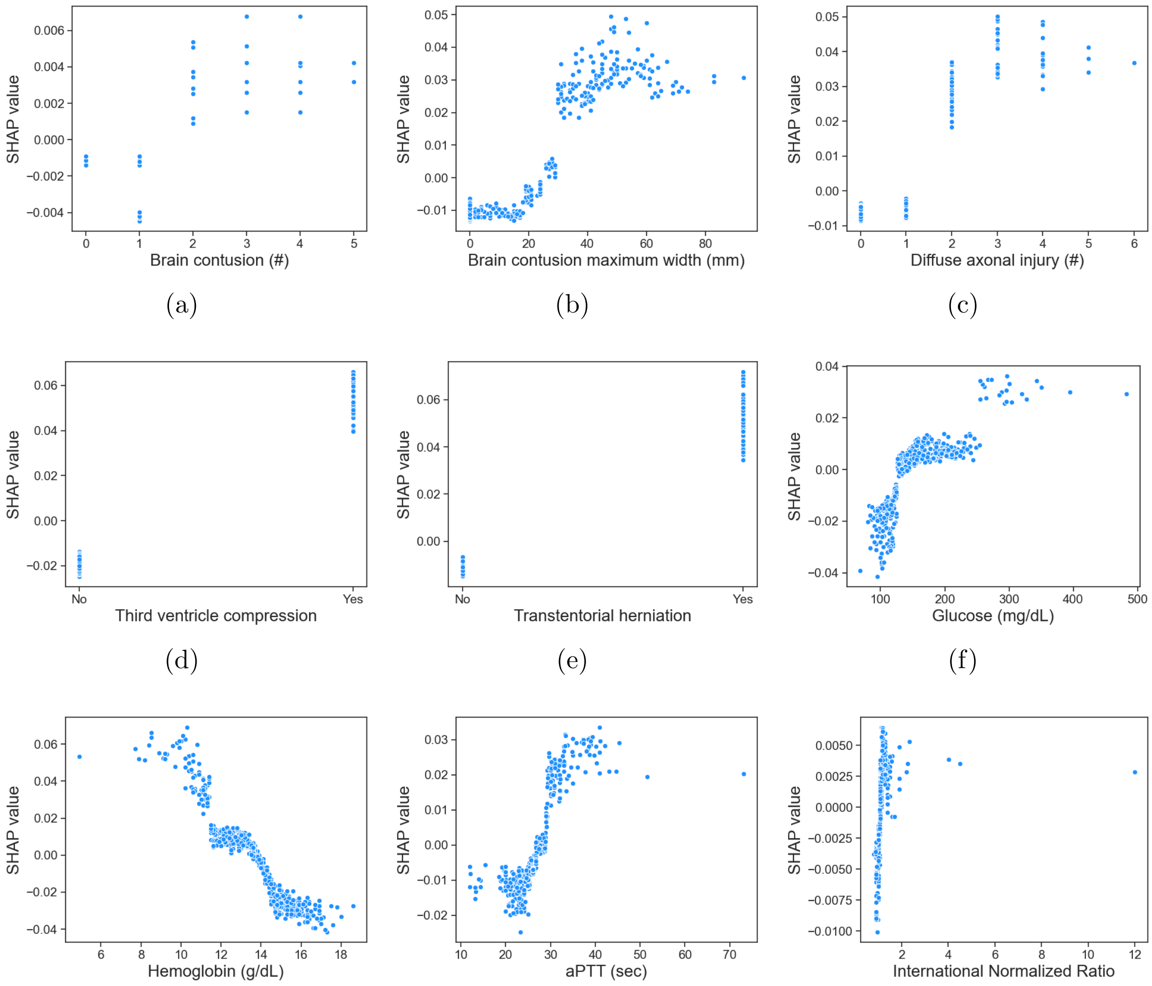
(e)

(f)

(g)

(h)

(i)



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure A.3: Detailed contribution of the 18 selected features to the predicted risk for TBI functional outcome (*Farzaneh et al.*, 2021b).

APPENDIX B

Additional Information for Chapter III

Table B.1: The comprehensive list of the utilized hand-crafted and deep features for subdural hematoma segmentation.

Feature Type	**Feature/Feature Class**	
Domain knowledge-driven	Age	
	Location-based	Radial distance
		Azimuth angle
		Elevation angle
		Distance to skull
	Histogram-based	Minimum
		Maximum
		Average
		Standard deviation
		Skewness
		Kurtosis
		Entropy
	Filtering-based	Gabor
		Laplacian of Gaussian
Data-driven	Deep features	

APPENDIX C

Additional Information for Chapter V

Liver Probability Estimation

For liver probability estimation two parameters of position and intensity were considered for each pixel of the image. In this section, generation of the position based probability atlas and intensity based probability atlas are discussed.

Position probability provides the probability of each pixel being liver based on its coordinates. In order to make the position probability model, in each slice of the annotated set, the value of pixels labeled as liver was set to 1, and the value of pixels not labeled as liver was set to 0. Then, for each slice in a CT scan, the corresponding slices in all subjects (i.e., registered images) were averaged. The resulting values would be from 0 to 1, stating the probability of pixel being liver based on its position in an image.

By having annotated liver regions on CT scans, a profile of liver pixel intensities was generated. The same profile was generated for background pixels (i.e., anything outside the liver boundary in the abdomen). To calculate the probability of being liver for a pixel with intensity I, we divided the number of pixels within the liver mask having the intensity I by the total number of pixels in both foreground and background having the same intensity I. This way the probability of pixel being liver based on its intensity was calculated for all possible intensity values of I. To calculate the probability of each pixel belonging to the liver, these two location- and intensity-based probabilities were multiplied.

Table C.1: The comprehensive list of the utilized hand-crafted features for kidney segmentation.

Feature Class	**Feature**
Location-based	X-coordinate
	Y-coordinate
	Z-coordinate
Histogram-based	Minimum
	Maximum
	Average
	Standard deviation
	Skewness
	Kurtosis
	Entropy
Filtering-based	Gabor
	Gaussian
	Laplacian of Gaussian

BIBLIOGRAPHY

BIBLIOGRAPHY

3DIRCAD (Accessed: February, 2021), 3dircad data set., https://www.ircad.fr/research/3d-ircadb-01.

AAST (Accessed: February, 2021), American association for surgery trauma., https://www.aast.org/resources-detail/injury-scoring-scale.

Achanta, R., A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk (2012), SLIC superpixels compared to state-of-the-art superpixel methods, *IEEE transactions on pattern analysis and machine intelligence*, *34*(11), 2274–2282.

Ahmad, M., D. Ai, G. Xie, S. F. Qadri, H. Song, Y. Huang, Y. Wang, and J. Yang (2019), Deep belief network modeling for automatic liver segmentation, *IEEE Access*, *7*, 20,585–20,595.

Ahmed, N., and J. J. Vernick (2011), Management of liver trauma in adults, *Journal of Emergencies, Trauma and Shock*, *4*(1), 114.

Arumugam, S., A. Al-Hassani, A. El-Menyar, H. Abdelrahman, A. Parchani, R. Peralta, A. Zarour, and H. Al-Thani (2015), Frequency, causes and pattern of abdominal trauma: a 4-year descriptive analysis, *Journal of emergencies, trauma, and shock*, *8*(4), 193.

Badger, S., R. Barclay, P. Campbell, D. Mole, and T. Diamond (2009), Management of liver trauma, *World journal of surgery*, *33*(12), 2522–2537.

Barber, C. B., D. P. Dobkin, and H. Huhdanpaa (1996), The quickhull algorithm for convex hulls, *ACM Transactions on Mathematical Software (TOMS)*, *22*(4), 469–483.

Bardera, A., I. Boada, M. Feixas, S. Remollo, G. Blasco, Y. Silva, and S. Pedraza (2009), Semi-automated method for brain hematoma and edema quantification using computed tomography, *Computerized medical imaging and graphics*, *33*(4), 304–311.

Barrie, J., S. Jamdar, M. Iniguez, O. Bouamra, T. Jenks, F. Lecky, and D. O'Reilly (2018), Improved outcomes for hepatic trauma in england and wales over a decade of trauma and hepatobiliary surgery centralisation, *European Journal of Trauma and Emergency Surgery*, *44*(1), 63–70.

Becker, K., A. Baxter, W. Cohen, H. Bybee, D. Tirschwell, D. Newell, H. Winn, and W. Longstreth (2001), Withdrawal of support in intracerebral hemorrhage may lead to self-fulfilling prophecies, *Neurology*, *56*(6), 766–772.

Bhadauria, H. S., A. Singh, and M. L. Dewal (2013), An integrated method for hemorrhage segmentation from brain CT imaging, *Computers & Electrical Engineering*, *39*(5), 1527–1536.

Bhandari, S., A. K. Kukreja, A. Lazar, A. Sim, and K. Wu (2020), Feature selection improves tree-based classification for wireless intrusion detection, in *Proceedings of the 3rd International Workshop on Systems and Network Telemetry and Analytics*, pp. 19–26.

Bi, Y., D. Xiang, Z. Ge, F. Li, C. Jia, and J. Song (2020), An interpretable prediction model for identifying n7-methylguanosine sites based on xgboost and shap, *Molecular Therapy-Nucleic Acids*.

Birnbaum, B. A., J. E. Jacobs, and P. Ramchandani (1996), Multiphasic renal ct: comparison of renal mass enhancement during the corticomedullary and nephrographic phases., *Radiology*, *200*(3), 753–758.

Bland, J. M., and D. Altman (1986), Statistical methods for assessing agreement between two methods of clinical measurement, *The lancet*, *327*(8476), 307–310.

Bland, J. M., and D. G. Altman (2010), Statistical methods for assessing agreement between two methods of clinical measurement, *International journal of nursing studies*, *47*(8), 931–936.

Buquicchio, G. L., G. Cuneo, S. Giannecchini, R. Palliola, M. Trinci, and V. Miele (2018), The follow-up of patients with abdominal injuries, in *Diagnostic Imaging in Polytrauma Patients*, pp. 509–532, Springer.

Bydder, G., R. Chapman, D. Harry, L. Bassan, S. Sherlock, and L. Kreel (1981), Computed tomography attenuation values in fatty liver., *The Journal of computed tomography*, *5*(1), 33.

Caruana, R., Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad (2015), Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission, in *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1721–1730.

Chan, T. (2007), Computer aided detection of small acute intracranial hemorrhage on computer tomography of brain, *Computerized Medical Imaging and Graphics*, *31*(4), 285–298.

Chan, T. F., and L. A. Vese (2001), Active contours without edges, *IEEE Transactions on image processing*, *10*(2), 266–277.

Chen, T., and C. Guestrin (2016), Xgboost: A scalable tree boosting system, in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794.

Chilamkurthy, S., R. Ghosh, S. Tanamala, M. Biviji, N. G. Campeau, V. K. Venugopal, V. Mahajan, P. Rao, and P. Warier (2018), Development and validation of deep learning algorithms for detection of critical findings in head CT scans, *arXiv preprint arXiv:1803.05854*.

Choy, G., et al. (2018), Current applications and future impact of machine learning in radiology, *Radiology*, *288*(2), 318–328.

Christ, P. F., et al. (2016), Automatic liver and lesion segmentation in ct using cascaded fully convolutional neural networks and 3d conditional random fields, in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 415–423, Springer.

Collaborators, M. C. T. (2008), Predicting outcome after traumatic brain injury: practical prognostic models based on large cohort of international patients, *BMJ: British Medical Journal*, *336*(7641), 425.

Committee on Medical Aspects of Automotive Safety (1971), Rating the severity of tissue damage. i. the abbreviated scale, *Jama*, *215*, 277–280.

Croce, M. A., T. C. Fabian, K. A. Kudsk, S. L. Baum, L. W. Payne, E. C. Mangiante, and L. G. Britt (1991), Aast organ injury scale: correlation of ct-graded liver injuries and operative findings., *The Journal of trauma*, *31*(6), 806–812.

Cuff, R. F., T. H. Cogbill, P. J. Lambert, C. E. Lucas, et al. (2000), Nonoperative management of blunt liver trauma: The value of follow-up abdominal computed tomography scans/discussion, *The American Surgeon*, *66*(4), 332.

Dayal, M., S. Gamanagatti, and A. Kumar (2013), Imaging in renal trauma, *World journal of radiology*, *5*(8), 275.

Deepika, A., and D. Shukla (2016), Prophesy in traumatic brain injury, *Journal of neurosciences in rural practice*, *7*(Suppl 1), S1.

Doklestić, K., et al. (2015), Surgical management of aast grades iii-v hepatic trauma by damage control surgery with perihepatic packing and definitive hepatic repair–single centre experience, *World Journal of Emergency Surgery*, *10*(1), 34.

Dreizin, D., et al. (2021), Added value of deep learning-based liver parenchymal ct volumetry for predicting major arterial injury after blunt hepatic trauma: a decision tree analysis, *Abdominal Radiology*, pp. 1–11.

Duan, R., M. Cao, Y. Wu, J. Huang, J. C. Denny, H. Xu, and Y. Chen (2016), An empirical study for impacts of measurement errors on ehr based association studies, in *AMIA Annual Symposium Proceedings*, vol. 2016, p. 1764, American Medical Informatics Association.

Elshawi, R., M. H. Al-Mallah, and S. Sakr (2019), On the interpretability of machine learning-based model for predicting hypertension, *BMC medical informatics and decision making*, *19*(1), 146.

Esteva, A., et al. (2021), Deep learning-enabled medical computer vision, *NPJ Digital Medicine*, *4*(1), 1–9.

Farzaneh, N., S. Samavi, S. R. Soroushmehr, H. Patel, S. Habbo-Gavin, D. P. Fessell, K. R. Ward, and K. Najarian (2016), Liver segmentation using location and intensity probabilistic atlases, in *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 6453–6456, IEEE.

Farzaneh, N., S. Habbo-Gavin, S. R. Soroushmehr, H. Patel, D. P. Fessell, K. R. Ward, and K. Najarian (2017a), Atlas based 3d liver segmentation using adaptive thresholding and superpixel approaches, in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1093–1097, IEEE.

Farzaneh, N., S. R. Soroushmehr, C. A. Williamson, C. Jiang, A. Srinivasan, J. R. Bapuraj, K. R. Ward, F. K. Korley, and K. Najarian (2017b), Automated subdural hematoma segmentation for traumatic brain injured (TBI) patients, in *Engineering in Medicine and Biology Society (EMBC), 2017 39th Annual International Conference of the IEEE*, pp. 3069–3072, IEEE.

Farzaneh, N., S. R. Soroushmehr, H. Patel, A. Wood, J. Gryak, D. Fessell, and K. Najarian (2018), Automated kidney segmentation for traumatic injured patients through ensemble learning and active contour modeling, in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 3418–3421, IEEE.

Farzaneh, N., C. A. Williamson, C. Jiang, A. Srinivasan, J. R. Bapuraj, J. Gryak, K. Najarian, and S. Soroushmehr (2020), Automated segmentation and severity analysis of subdural hematoma for patients with traumatic brain injuries, *Diagnostics*, *10*(10), 773.

Farzaneh, N., E. B. Stein, S. Soroushmehr, J. Gryak, K. Najarian, and S. Soroushmehr (2021a), A deep learning framework for automated detection and quantitative assessment of liver trauma (in review), *BMC Medical Imaging*.

Farzaneh, N., C. A. Williamson, J. Gryak, and K. Najarian (2021b), A hierarchical expert-guided machine learning framework for clinical decision support systems:an application to traumatic brain injury prognostication (accepted), *NPJ digital medicine*.

Faul, M., L. Xu, M. M. Wald, and V. G. Coronado (2010), Traumatic brain injury in the united states: Emergency department visits, hospitalizations and deaths 2002-2006., *Atlanta, GA: CDC, National Center for Injury Prevention and Control*.

Fogel, A. L., and J. C. Kvedar (2018), Artificial intelligence powers digital medicine, *NPJ digital medicine*, *1*(1), 1–4.

Gennarelli, T. A., and E. Wodzin (2008), *Abbreviated injury scale 2005: update 2008*, Russ Reeder.

Geurts, M., M. R. Macleod, G. J. van Thiel, J. van Gijn, L. J. Kappelle, and H. B. van der Worp (2014), End-of-life decisions in patients with severe acute brain injury, *The Lancet Neurology*, *13*(5), 515–524.

Goldstein, F. C., A. F. Caveney, V. S. Hertzberg, R. Silbergleit, S. D. Yeatts, Y. Y. Palesch, H. S. Levin, D. W. Wright, and N. Investigators (2017), Very early administration of progesterone does not improve neuropsychological outcomes in subjects with moderate to severe traumatic brain injury, *Journal of neurotrauma*, *34*(1), 115–120.

Grewal, M., M. M. Srivastava, P. Kumar, and S. Varadarajan (2018), Radnet: Radiologist level accuracy using deep learning for hemorrhage detection in CT scans, in *Biomedical Imaging (ISBI 2018), 2018 IEEE 15th International Symposium on*, pp. 281–284, IEEE.

Gwinn, E. C., and P. K. Park (2020), Blunt abdominal trauma, in *Evidence-Based Critical Care*, pp. 651–658, Springer.

Heimann, T., et al. (2009), Comparison and evaluation of methods for liver segmentation from ct datasets, *IEEE transactions on medical imaging*, *28*(8), 1251–1265.

Hemphill III, J. C., and D. B. White (2009), Clinical nihilism in neuroemergencies, *Emergency medicine clinics of North America*, *27*(1), 27–37.

Holzinger, A., C. Biemann, C. S. Pattichis, and D. B. Kell (2017), What do we need to build explainable ai systems for the medical domain?, *arXiv preprint arXiv:1712.09923*.

Hukkelhoven, C. W., E. W. Steyerberg, J. D. F. Habbema, E. Farace, A. Marmarou, G. D. Murray, L. F. Marshall, and A. I. Maas (2005), Predicting outcome after traumatic brain injury: development and validation of a prognostic score based on admission characteristics, *Journal of neurotrauma*, *22*(10), 1025–1039.

Hutchinson, P. J., et al. (2016), Trial of decompressive craniectomy for traumatic intracranial hypertension, *N Engl J Med*, *375*, 1119–1130.

Jain, A. K., and F. Farrokhnia (1991), Unsupervised texture segmentation using gabor filters, *Pattern recognition*, *24*(12), 1167–1186.

Janizek, J. D., S. Celik, and S.-I. Lee (2018), Explainable machine learning prediction of synergistic drug combinations for precision cancer medicine, *bioRxiv*, p. 331769.

Jansen, J., S. Yule, and M. Loudon (2008), Investigation of blunt abdominal trauma, *British Medical Journal*, *7650*, 938.

Joseph, B., et al. (2014), The significance of platelet count in traumatic brain injury patients on antiplatelet therapy, *Journal of Trauma and Acute Care Surgery*, *77*(3), 417–421.

Junior, J. R., L. C. Welling, M. Schafranski, L. T. Yeng, R. R. do Prado, E. Koterba, A. F. de Andrade, M. J. Teixeira, and E. G. Figueiredo (2017), Prognostic model for patients with traumatic brain injuries and abnormal computed tomography scans, *Journal of clinical neuroscience*.

Kawashima, A., C. M. Sandler, F. M. Corl, O. C. West, E. P. Tamm, E. K. Fishman, and S. M. Goldman (2001), Imaging of renal trauma: a comprehensive review, *Radiographics*, *21*(3), 557–574.

Kelly, C. J., A. Karthikesalingam, M. Suleyman, G. Corrado, and D. King (2019), Key challenges for delivering clinical impact with artificial intelligence, *BMC medicine*, *17*(1), 195.

Khalifa, F., A. Soliman, A. C. Dwyer, G. Gimel'farb, and A. El-Baz (2016), A random forest-based framework for 3D kidney segmentation from dynamic contrast-enhanced CT images, in *Image Processing (ICIP), 2016 IEEE International Conference on*, pp. 3399–3403, IEEE.

Khalifa, F., A. Soliman, A. Elmaghraby, G. Gimel'farb, and A. El-Baz (2017), 3D kidney segmentation from abdominal images using spatial-appearance models, *Computational and mathematical methods in medicine*, *Volume 2017*.

Lebre, M.-A., A. Vacavant, M. Grand-Brochier, H. Rositi, A. Abergel, P. Chabrot, and B. Magnin (2019a), Automatic segmentation methods for liver and hepatic vessels from ct and mri volumes, applied to the couinaud scheme, *Computers in biology and medicine*, *110*, 42–51.

Lebre, M.-A., A. Vacavant, M. Grand-Brochier, H. Rositi, R. Strand, H. Rosier, A. Abergel, P. Chabrot, and B. Magnin (2019b), A robust multi-variability model based liver segmentation algorithm for ct-scan and mri modalities, *Computerized Medical Imaging and Graphics*, *76*, 101,635.

Liao, C.-C., F. Xiao, J.-M. Wong, and I.-J. Chiang (2009), A multiresolution binary level set method and its application to intracranial hematoma segmentation, *Computerized Medical Imaging and Graphics*, *33*(6), 423–430.

Liao, C.-C., F. Xiao, J.-M. Wong, and I.-J. Chiang (2010), Computer-aided diagnosis of intracranial hematoma with brain deformation on computed tomography, *Computerized Medical Imaging and Graphics*, *34*(7), 563–571.

Lin, D.-T., C.-C. Lei, and S.-W. Hung (2006), Computer-aided kidney segmentation on abdominal CT images, *IEEE transactions on information technology in biomedicine*, *10*(1), 59–65.

Liu, B., Q. Yuan, Z. Liu, X. Li, and X. Yin (2008), Automatic segmentation of intracranial hematoma and volume measurement, in *Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE*, pp. 1214–1217, IEEE.

Liu, J., S. Wang, M. G. Linguraru, J. Yao, and R. M. Summers (2015), Computer-aided detection of exophytic renal lesions on non-contrast CT images, *Medical image analysis*, *19*(1), 15–29.

Lu, F., F. Wu, P. Hu, Z. Peng, and D. Kong (2017), Automatic 3d liver location and segmentation via convolutional neural network and graph cut, *International journal of computer assisted radiology and surgery*, *12*(2), 171–182.

Lundberg, S. M., and S.-I. Lee (2017), A unified approach to interpreting model predictions, in *Advances in neural information processing systems*, pp. 4765–4774.

Lundberg, S. M., G. G. Erion, and S.-I. Lee (2018), Consistent individualized feature attribution for tree ensembles, *arXiv preprint arXiv:1802.03888.*

Ma, S., and R. Tourani (2020), Predictive and causal implications of using shapley value for model interpretation, in *Proceedings of the 2020 KDD Workshop on Causal Discovery*, pp. 23–38, PMLR.

Maas, A. I., C. W. Hukkelhoven, L. F. Marshall, and E. W. Steyerberg (2005), Prediction of outcome in traumatic brain injury with computed tomographic characteristics: a comparison between the computed tomographic classification and combinations of computed tomographic predictors, *Neurosurgery*, *57*(6), 1173–1182.

Maas, A. I., N. Stocchetti, and R. Bullock (2008), Moderate and severe traumatic brain injury in adults, *The Lancet Neurology*, *7*(8), 728–741.

Maegele, M. (2013), Coagulopathy after traumatic brain injury: incidence, pathogenesis, and treatment options, *Transfusion*, *53*, 28S–37S.

Majdan, M., A. Brazinova, M. Rusnak, and J. Leitgeb (2017), Outcome prediction after traumatic brain injury: Comparison of the performance of routinely used severity scores and multivariable prognostic models, *Journal of neurosciences in rural practice*, *8*(1), 20.

Marshall, L., S. B. Marshall, M. R. Klauber, M. B. C. Van, H. Eisenberg, J. Jane, T. Luerssen, A. Marmarou, and M. Foulkes (1992), The diagnosis of head injury requires a classification based on computed axial tomography., *Journal of neurotrauma*, *9*, S287–92.

Marshall, L. F., S. B. Marshall, M. R. Klauber, M. van Berkum Clark, H. M. Eisenberg, J. A. Jane, T. G. Luerssen, A. Marmarou, and M. A. Foulkes (1991), A new classification of head injury based on computerized tomography, *Journal of neurosurgery*, *75*(Supplement), S14–S20.

Matsuo, K., H. Aihara, T. Nakai, A. Morishita, Y. Tohma, and E. Kohmura (2020), Machine learning to predict in-hospital morbidity and mortality after traumatic brain injury, *Journal of Neurotrauma*, *37*(1), 202–210.

McMillan, T., L. Wilson, J. Ponsford, H. Levin, G. Teasdale, and M. Bond (2016), The glasgow outcome scale—40 years of application and refinement, *Nature Reviews Neurology*, *12*(8), 477–485.

MicroDicom (Accessed: May 1, 2020), Microdicom dicom viewer., https://www.microdicom.com.

Moore, N., P. Brennan, and J. Baillie (2013), Wide variation and systematic bias in expert clinicians' perceptions of prognosis following brain injury, *British journal of neurosurgery*, *27*(3), 340–343.

Moppett, I. K. (2007), Traumatic brain injury: assessment, resuscitation and early management, *British Journal of Anaesthesia*, *99*(1), 18–31.

Nellensteijn, D., H. Ten Duis, J. Oldenziel, W. Polak, and J. Hulscher (2009), Only moderate intra-and inter-observer agreement between radiologists and surgeons when grading blunt paediatric hepatic injury on ct scan, *European journal of pediatric surgery*, *19*(06), 392–394.

Ogura, K., T. Goto, T. Shirakawa, T. Sonoo, H. Nakano, and K. Nakamura (2020), Development of prediction model for trauma assessment using electronic medical records, *medRxiv*.

Okada, T., R. Shimada, M. Hori, M. Nakamoto, Y.-W. Chen, H. Nakamura, and Y. Sato (2008), Automated segmentation of the liver from 3d ct images using probabilistic atlas and multilevel statistical shape model, *Academic radiology*, *15*(11), 1390–1403.

Park, S., J. Han, T. Kim, and B. Choi (2001), Three-dimensional spiral CT cholangiography with minimum intensity projection in patients with suspected obstructive biliary disease: comparison with percutaneous transhepatic cholangiography, *Abdominal imaging*, *26*(3), 281–286.

Paulus, J. K., and D. M. Kent (2017), Race and ethnicity: A part of the equation for personalized clinical decision making?, *Circulation: Cardiovascular Quality and Outcomes*, *10*(7), e003,823.

Paulus, J. K., and D. M. Kent (2020), Predictably unequal: understanding and addressing concerns that algorithmic clinical prediction may increase health disparities, *NPJ digital medicine*, *3*(1), 1–8.

Piper, G. L., and A. B. Peitzman (2010), Current management of hepatic trauma, *Surgical Clinics*, *90*(4), 775–785.

Powers, W. F., L. N. Beard, A. Adams, C. A. Kotwall, T. V. Clancy, and W. W. Hope (2012), Solid organ injury grading in trauma: accuracy of grading by surgical residents, *The American Surgeon*, *78*(8), 834–836.

Rafiei, S., N. Karimi, B. Mirmahboub, K. Najarian, B. Felfeliyan, S. Samavi, and S. R. Soroushmehr (2019), Liver segmentation in abdominal ct images using probabilistic atlas and adaptive 3d region growing, in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 6310–6313, IEEE.

Rajkomar, A., J. Dean, and I. Kohane (2019), Machine learning in medicine, *New England Journal of Medicine*, *380*(14), 1347–1358.

Rau, C.-S., P.-J. Kuo, P.-C. Chien, C.-Y. Huang, H.-Y. Hsieh, and C.-H. Hsieh (2018), Mortality prediction in patients with isolated moderate and severe traumatic brain injury using machine learning models, *PloS one*, *13*(11), e0207,192.

Rhee, P., B. Joseph, V. Pandit, H. Aziz, G. Vercruysse, N. Kulvatunyou, and R. S. Friese (2014), Increasing trauma deaths in the united states, *Annals of surgery*, *260*(1), 13–21.

Rizoli, S., et al. (2016), Early prediction of outcome after severe traumatic brain injury: a simple and practical model, *BMC emergency medicine*, *16*(1), 32.

Ronneberger, O., P. Fischer, and T. Brox (2015), U-net: Convolutional networks for biomedical image segmentation, in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2015 18th International Conference on*, pp. 234–241, Springer.

Saatman, K. E., A.-C. Duhaime, R. Bullock, A. I. Maas, A. Valadka, and G. T. Manley (2008), Classification of traumatic brain injury for targeted therapies, *Journal of neurotrauma*, *25*(7), 719–738.

Sass, D. A., P. Chang, and K. B. Chopra (2005), Nonalcoholic fatty liver disease: a clinical review, *Digestive diseases and sciences*, *50*(1), 171.

Shahangian, B., and H. Pourghassem (2013), Automatic brain hemorrhage segmentation and classification in CT scan images, in *Machine Vision and Image Processing (MVIP), 2013 8th Iranian Conference on*, pp. 467–471, IEEE.

Shahangian, B., and H. Pourghassem (2016), Automatic brain hemorrhage segmentation and classification algorithm based on weighted grayscale histogram feature in a hierarchical classification structure, *Biocybernetics and Biomedical Engineering*, *36*(1), 217–232.

Shi, C., Y. Cheng, J. Wang, Y. Wang, K. Mori, and S. Tamura (2017), Low-rank and sparse decomposition based shape model and probabilistic atlas for automatic pathological organ segmentation, *Medical image analysis*, *38*, 30–49.

Skalski, A., K. Heryan, J. Jakubowski, and T. Drewniak (2017), Kidney segmentation in CT data using hybrid level-set method with ellipsoidal shape constraints, *Metrology and Measurement Systems*, *24*(1), 101–112.

Sofka, M., D. Wu, M. Sühling, D. Liu, C. Tietjen, G. Soza, and S. K. Zhou (2011), Automatic contrast phase estimation in CT volumes, in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 166–174, Springer.

Soler, L., et al. (2001), Fully automatic anatomical, pathological, and functional segmentation from CT scans for hepatic surgery, *Computer Aided Surgery*, *6*(3), 131–142.

Stenberg, M., L.-O. D. Koskinen, P. Jonasson, R. Levi, and B.-M. Stålnacke (2017), Computed tomography and clinical outcome in patients with severe traumatic brain injury, *Brain injury*, *31*(3), 351–358.

Steyerberg, E. W., et al. (2008), Predicting outcome after traumatic brain injury: development and international validation of prognostic scores based on admission characteristics, *PLoS medicine*, *5*(8), e165.

Taghavi, S., and R. Askari (2019), Liver trauma, in *StatPearls [Internet]*, StatPearls Publishing.

Teasdale, G., and B. Jennett (1974), Assessment of coma and impaired consciousness: a practical scale, *The Lancet*, *304*(7872), 81–84.

Teasdale, G., and B. Jennett (1976), Assessment and prognosis of coma after head injury, *Acta neurochirurgica*, *34*(1-4), 45–55.

Vellido, A. (2019), The importance of interpretability and visualization in machine learning for applications in medicine and health care, *Neural Computing and Applications*, pp. 1–15.

Vorontsov, E., G. Chartrand, A. Tang, C. Pal, and S. Kadoury (2017), Liver lesion segmentation informed by joint liver segmentation, *arXiv preprint arXiv:1707.07734*.

Wells, M. L., E. R. Fenstad, J. T. Poterucha, D. M. Hough, P. M. Young, P. A. Araoz, R. L. Ehman, and S. K. Venkatesh (2016), Imaging findings of congestive hepatopathy, *Radiographics*, *36*(4), 1024–1037.

Whitaker, R. T. (1998), A level-set approach to 3d reconstruction from range data, *International journal of computer vision*, *29*(3), 203–231.

Wijdicks, E. F., W. R. Bamlet, B. V. Maramattom, E. M. Manno, and R. L. McClelland (2005), Validation of a new coma scale: the four score, *Annals of neurology*, *58*(4), 585–593.

Williamson, C. A., and V. Rajajee (2018), Intracerebral hemorrhage prognosis, in *Intracerebral Hemorrhage Therapeutics*, pp. 95–105, Springer.

Wintermark, M., et al. (2015), Imaging evidence and recommendations for traumatic brain injury: conventional neuroimaging techniques, *Journal of the American College of Radiology*, *12*(2), e1–e14.

Wolz, R., C. Chu, K. Misawa, K. Mori, and D. Rueckert (2012), Multi-organ abdominal CT segmentation using hierarchically weighted subject-specific atlases, *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2012*, pp. 10–17.

Wolz, R., C. Chu, K. Misawa, M. Fujiwara, K. Mori, and D. Rueckert (2013), Automated abdominal multi-organ segmentation with subject-specific atlas generation, *IEEE transactions on medical imaging*, *32*(9), 1723–1730.

Wright, D. W., et al. (2014), Very early administration of progesterone for acute traumatic brain injury, *New England Journal of Medicine*, *371*(26), 2457–2466.

Yuh, E. L., A. D. Gean, G. T. Manley, A. L. Callen, and M. Wintermark (2008), Computer-aided assessment of head computed tomography (CT) studies in patients with suspected traumatic brain injury, *Journal of neurotrauma*, *25*(10), 1163–1172.