Convolutional Neural Networks for the Automated Segmentation and Recurrence Risk Prediction of Surgically Resected Lung Tumors

The Harvard community has made this article openly available. Please share how this access benefits you. Your story matters

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Citable link</td>
<td><a href="https://nrs.harvard.edu/URN-3:HUL.INSTREPOS:37364751">https://nrs.harvard.edu/URN-3:HUL.INSTREPOS:37364751</a></td>
</tr>
<tr>
<td>Terms of Use</td>
<td>This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at <a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA">http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA</a></td>
</tr>
</tbody>
</table>
Convolutional Neural Networks for the Automated Segmentation and Recurrence Risk Prediction of Surgically Resected Lung Tumors

A thesis presented by

Marguerite Basta

to

the Faculty of the
Harvard John A. Paulson School of Engineering and Applied Sciences
in partial fulfillment of the requirements for
the Joint Bachelor of Arts degree
in Electrical Engineering and Computer Science

Faculty Advisor: Flavio P. Calmon

Harvard University
Cambridge, MA
March 27, 2020
In submitting this thesis to the Harvard John A. Paulson School of Engineering and Applied Sciences in partial fulfillment of the requirements for the degree with honors of Bachelor of Arts, I affirm my awareness of the standards of the Harvard College Honor Code.

Name: Marguerite Basta

Signature: 

The Harvard College Honor Code Members of the Harvard College community commit themselves to producing academic work of integrity; that is, work that adheres to the scholarly and intellectual standards of accurate attribution of sources, appropriate collection and use of data, and transparent acknowledgement of the contribution of others to their ideas, discoveries, interpretations, and conclusions. Cheating on exams or problem sets, plagiarizing or misrepresenting the ideas or language of someone else as one’s own, falsifying data, or any other instance of academic dishonesty violates the standards of our community, as well as the standards of the wider world of learning and affairs.
To my mom. For showing me anything is possible.
Acknowledgements

Without the support and generosity of others, this thesis would not have been possible.

I first want to thank my thesis advisor, Flavio Calmon. Ever since I had privilege of being his student in my sophomore year, Flavio has gone above and beyond to support my academic endeavors. His encouragement has given me the confidence and drive I needed in my pursuits as an undergraduate. Flavio, thank you for believing in me.

I would also like to thank the others who have been generous enough to dedicate their time to this thesis. I have been extremely fortunate to collaborate with Sarafaz Hussein. Sarfaraz was kind enough to answer an out of the blue email and give his time to a student he had never met. His experience in deep learning within medical imaging has been an invaluable resource throughout the creation of this thesis. I would also like to thank Hsiang Hsu. From the start, Hsiang has taken time away from his own research to meet regularly with me and Flavio. His insight and critique has raised the standard of my work. Finally, I want to express my appreciation for the other readers of my thesis, Todd Zickler and Finale Doshi-Velez. I am extremely grateful for their willingness to support and evaluate my research.

Lastly, I want to thank my family, my friends, and my other advisors here at Harvard. They have all enabled me, pushed me, and believed in me, and they have done so unwaveringly. Thank you all for showing me what can be can be done when you set your mind to something. I would not be here otherwise.
Abstract

Lung cancer is the leading cause of cancer related mortality by a significant margin. While new technologies, such as image segmentation, have been paramount to improved detection and earlier diagnoses, there are still significant challenges in treating the disease. In particular, despite an increased number of curative resections, many postoperative patients still develop recurrent lesions. Consequently, there is a significant need for prognostic tools that can more accurately predict a patient’s risk for recurrence.

In this thesis, we explore the use of convolutional neural networks (CNNs) for the segmentation and recurrence risk prediction of lung tumors that are present in pre-operative computed tomography (CT) images. First, expanding upon recent progress in medical image segmentation, a residual U-Net is used to localize and geometrically characterize each nodule. Then, the identified tumors are passed to a second CNN for recurrence risk prediction. The system’s final results are produced with a random forest classifier that synthesizes the predictions of the second network with clinical and geometric attributes. Our proposed framework demonstrates that first, automated nodule segmentation methods can generalize to enable pipelines for a wide range of multitask systems and second, that deep learning and image processing have the potential to improve current prognostic tools. To the best of our knowledge, our proposed framework is the first fully automated segmentation and recurrence risk prediction system to be implemented.
# Table of Contents

List of Tables ix

List of Figures x

1 Introduction 1
   1.1 Lung Cancer and Recurrence ................................. 1
   1.2 Current State of Computer Aided Diagnostics .................. 3
   1.3 Current State of Recurrence Risk Prediction .................. 5
   1.4 Methods ......................................................... 7
   1.5 Challenges ...................................................... 11
   1.6 Significance ....................................................... 12
   1.7 Thesis Structure ............................................... 14

2 Materials and Preprocessing 15
   2.1 Datasets .......................................................... 15
   2.2 Data Cleaning and Feature Engineering ........................... 18
   2.3 Preprocessing ..................................................... 20
   2.4 Segmentation Target Generation ................................. 22

3 Lung Nodule Segmentation 24
   3.1 Background on Lung Nodule Segmentation ...................... 25
      3.1.1 Overview of Image Segmentation and its Medical Applications 25
      3.1.2 Review of Image Segmentation Methods ....................... 27
      3.1.3 Complexity of Lung Nodule Segmentation .................... 30
   3.2 Related Work ..................................................... 34
      3.2.1 U-Net .......................................................... 35

vi
3.2.2 Residual U-Net ................................................. 36
3.3 Proposed Method ............................................. 37
  3.3.1 Proposed Network Architecture ......................... 37
  3.3.2 Training ................................................. 39
  3.3.3 Post-Processing ......................................... 41
3.4 Results and Experiments ................................... 43
  3.4.1 Evaluation Metrics ..................................... 43
  3.4.2 Results ................................................. 45
  3.4.3 Further Comparisons ................................... 46
  3.4.4 SPIE-AAPM Testing ..................................... 47
3.5 Discussion .................................................. 48

4 Recurrence Prediction ........................................ 50
  4.1 Background on Recurrence Prediction ..................... 51
    4.1.1 Overview of Recurrence in Lung Cancer .............. 51
    4.1.2 The TNM System ...................................... 52
  4.2 Related Work ............................................... 53
  4.3 Proposed Method ......................................... 55
    4.3.1 ROI Generation ....................................... 56
    4.3.2 Proposed Network Architecture ........................ 57
    4.3.3 Training ............................................. 58
    4.3.4 Integrating Clinical and Geometric Data With a Random Forest Classifier .................................. 59
  4.4 Results and Experiments ................................... 62
    4.4.1 Evaluation Metrics ................................... 62
    4.4.2 Results .............................................. 63
  4.5 Discussion ................................................ 68

5 Conclusion .................................................... 70
  5.1 Overview .................................................. 70
  5.2 Contributions ............................................ 71
  5.3 Limitations and Future Work .............................. 72
  5.4 Concluding Remarks ...................................... 74
## List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Dataset Summary</td>
<td>20</td>
</tr>
<tr>
<td>3.1</td>
<td>Segmentation Results Before Post-Processing</td>
<td>46</td>
</tr>
<tr>
<td>3.2</td>
<td>Segmentation Results After Post-Processing</td>
<td>46</td>
</tr>
<tr>
<td>3.3</td>
<td>Segmentation Results for Percent of Primary Nodule Included in ROI</td>
<td>46</td>
</tr>
<tr>
<td>3.4</td>
<td>Comparison With Other Segmentation Models</td>
<td>47</td>
</tr>
<tr>
<td>4.1</td>
<td>Neural Network Overall Recurrence Prediction Performance</td>
<td>64</td>
</tr>
<tr>
<td>4.2</td>
<td>Neural Network Recurrence Prediction Performance by Cancer Stage</td>
<td>64</td>
</tr>
<tr>
<td>4.3</td>
<td>Random Forest Overall Recurrence Prediction Performance</td>
<td>66</td>
</tr>
<tr>
<td>4.4</td>
<td>Random Forest Recurrence Prediction Performance by Cancer Stage</td>
<td>66</td>
</tr>
</tbody>
</table>
## List of Figures

1.1 World Lung Cancer Deaths ........................................... 2
1.2 LIDC-IDRI Image Example ........................................... 5
1.3 Inter-Reader Variability Between Radiologist Annotations ...... 6
1.4 3D ROI Around Lung Nodule ......................................... 8
1.5 Proposed Nodule Segmentation and Recurrence Prediction Framework ........................................... 9
1.6 Residual U-Net Architecture for Segmentation .................... 10
1.7 3D CNN Architecture for Recurrence Risk Prediction .......... 11
2.1 Preprocessing Pipeline ............................................... 21
2.2 Image with Challenging Lung Field Segmentation ............... 22
2.3 Binary Target Generation for Segmentation ...................... 23
3.1 Segmentation Pipeline ............................................... 24
3.2 Example of Image Segmentation .................................... 25
3.3 Example of Thresholding Methods for Segmentation ............ 28
3.4 Example of Region Growing for Segmentation ................... 29
3.5 Lung CT Regions and Histograms ................................ 31
3.6 Lung Nodule Attenuation ........................................... 32
3.7 Lung Nodule Embedded in the Lung Wall ......................... 33
3.8 Similarities Between Lung Nodules and Other Anatomical Structures Within Lungs ........................................... 34
3.9 The Original U-Net Architecture .................................. 35
3.10 Residual Unit Architecture ....................................... 36
3.11 Architecture of the Proposed Residual U-Net for Segmentation ........................................... 38
3.12 Segmentation Network Data Augmentation ....................... 41
3.13 Post-Processing .................................................... 43
Chapter 1

Introduction

In this thesis, we explore the use of convolutional neural networks (CNNs) for the segmentation and recurrence risk prediction of lung tumors that are present in preoperative computed tomography (CT) images. By taking advantage of recent progress in image segmentation, we are able to implement a fully automated system that utilizes a novel, yet unannotated dataset. Towards the end of this chapter, we give an outline of our approach and its contributions. First however, we begin by providing background on lung cancer and the pressing challenges that are posed by disease recurrence.

1.1 Lung Cancer and Recurrence

Lung cancer causes significantly more deaths each year than any other form of cancer. In fact, even though there has been significant progress in the diagnosis and treatment of the disease, it is still projected to remain the leading cause of cancer related mortality through 2030 [1] (see Figure 1.1). While improved medical technology can enable earlier detection and therefore increase the number of curative tumor resections, over 30% of postoperative patients will still develop recurrent lesions [2]. On average, these patients do not survive more than a year [3]. Consequently, disease
Figure 1.1: World lung cancer deaths. Total number of world deaths from 1990-2017 caused by a range of different cancer types. Tracheal, bronchus, and lung cancers consistently remain the most deadly causing 1.9 million deaths in 2017. The data for the plot was collected by the Institute of Health Metrics and Evaluation (IHME) as part of the Global Burden of Disease (GBD) project and was organized by [5].

recurrence contributes significantly to lung cancer’s startlingly low 5 year survival rate of 4-17% depending on stage and geographic location [4].

With the rise of precision medicine and continued development of new treatment options however, there has been an increased push for better and more individualized recurrence prognostication tools. The failure to accurately evaluate risk for recurrence in the past is largely due to the inadequacy of current methods, particularly the Tumor, Node, Metastasis (TNM) staging system.\(^1\) The system is the medical standard for defining the extent and spread of cancer. Despite the fact that it lacks the granularity required to accurately predict risk of recurrence on an individual ba-

\(^1\)T (tumor) refers to the extent of the primary tumor. N (node) refers to the extent of cancer that has spread to nearby lymph nodes. M (metastasis) refers to whether the cancer has spread to distant parts of the body.
sis, in clinical settings, it is still considered the best tool for the task [6]. Nonetheless, new medical technologies are continuing to enable the development of potentially transformative approaches.

1.2 Current State of Computer Aided Diagnostics

Recent advances in machine learning and image processing have led to a dramatic increase in the efficacy of computer aided detection (CAD) and computer aided diagnosis (CADx). In particular, one area that has shown significant promise is radiomics — the emerging field of research that aims to extract high-dimensional data from radiographic medical images. Deep learning techniques like CNNs and their variants have enabled breakthroughs in cellular segmentation [7], nodule segmentation [8–12], malignancy classification [13, 14], and a variety of other medical tasks [15]. For lung cancer in particular, radiomics has exhibited significant potential to improve diagnostic technology [16, 17]. CT scanning has become the predominant form of screening for the disease and gives practitioners regular access to high quality images. Furthermore, the recent generation of extensive imaging datasets has enabled state of the art research [18, 19].

Lung nodule segmentation is one area that has progressed significantly in recent years. The accelerated development of these methods can largely be attributed to the 2011 release of the LIDC-IDRI dataset [18]. Published by the Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI), the dataset is the first large-scale lung cancer imaging collection that is both publically available and analyzed by multiple radiologists (see Figure 1.2). In addition to annotated nodules, the dataset includes radiologists’ malignancy scores. As a result, it has become

\[^2\]For each nodule, each radiologist provides a malignancy score on a scale from 1-5. A score of 1 corresponds to high confidence that the nodule is benign. A score of 5 corresponds to high confidence that the nodule is malignant.
particularly popular for the development of not just tumor segmentation algorithms
[8–12, 20], but also malignancy classification [13, 14], and joint learning [9] algorithms. The prevalence of the dataset has streamlined the development process for these tasks and established clear benchmarks for success.

Despite their exhibited potential, few of the algorithms trained on LIDC-IDRI have been expanded to other datasets. Therefore, the ability of these techniques to generalize is brought into question. By using LIDC-IDRI exclusively, the scope of possible tasks is limited. Since the only clinical information provided is radiologists’ malignancy scores, a model for any other diagnostic classification problem, like recurrence risk prediction, requires a different dataset. Alas, very few datasets have nodule annotations that are comparable to those of LIDC-IDRI. For many algorithms this is a significant impediment as these annotations are used to localize the tumors and reduce the problem space [9–14, 20]. Even popular segmentation algorithms only consider a cropped region around the nodule as a starting point [9–12, 20]. As a result, these approaches cannot be generalized to datasets where nodule contours or center of masses are not provided. Unfortunately, few medical image collections contain this level of localization data. The most common information available to locate the tumor, if any, is an axial slice number corresponding to a 2D cross section within the 3D series where the tumor is present. Nonetheless, a potential alternative is to use LIDC-IDRI trained segmentation algorithms (those that segment the entire image) to localize and geometrically characterize tumors from datasets without annotations. Doing so would enable a much wider range of classification tasks, including recurrence risk prediction, to be tackled with datasets that lack annotations but contain more clinical data. However, prior to this study, this approach has been left mostly untested. Additionally, the use of radiomics to predict lung cancer recurrence is still a relatively novel area.
1.3 Current State of Recurrence Risk Prediction

The search for a better prognostic tool for lung cancer recurrence has consumed researchers across domains for decades. Three common approaches are modeling clinical risk factors [21–23], using histology and immunohistochemistry (IHC) to examine protein expressions [24, 25], and gene expression-based analysis [26–28]. While several studies within these fields have exhibited promise, significant concerns have also been raised about these types of approaches [23, 29, 30]. We include a more in depth discussion of these methods and the concerns around them in Chapter 4. Despite these setbacks, novel approaches continue to be developed.

Recurrence prediction research has recently found its way into image processing and deep learning via radiomics. Multiple studies have found significant evidence that radiographic data can capture information related to pathophysiologies, histologies, and even gene expressions [30–33]. These associations indicate that imaging based
models might be able to integrate multiple sources of information that are potentially associated with recurrence. The few studies that have attempted to build these models have had promising results. For example, Koo et al. [34] present a model that correlates tumor features extracted from preoperative CT scans with recurrence. Additionally, Honsy et al. [15] propose a 3D CNN model to stratify lung cancer patients into two year mortality risk groups.

Despite the promise of these studies, both rely on radiologists to localize the tumors by hand and do not have a fully automated pipelines. As a result, extending methodologies in [15, 34] to unannotated datasets is intractable if a radiologist cannot be accessed. Even if one can be accessed, there are still significant barriers to acquiring their annotations. For large datasets, shear labor intensity can make contouring every single image impossible within a reasonable time frame. This is a particularly challenging issue in the case of recurrence prediction. As it is, it is rare to find a dataset that has enough clinical information to truly determine a patient’s recurrence.

Figure 1.3: Inter-reader variability between radiologist annotations. An example of possible inter-reader variability in the manual segmentation of tumors. The image above shows two different radiologists’ delineations of the same tumor [18].
status (as opposed to just survival time). In fact, to the best of our knowledge, there are no published lung imaging datasets that contain both the information to determine recurrence status and annotations from radiologists. Furthermore, not having automated segmentation has other consequences. In fact, reliance on the manual contours is an example of one of the largest challenges facing radiomics today — the lack of standardized practices [35]. As is evident in Figure 1.3, there is often high inter-reader variability involved in manual contouring. This variability can lead to results that are difficult to generalize or reproduce [35]. Therefore, demonstrating the reliability of automated segmentation methods across datasets and classification tasks is a significant step towards standardizing radiomics methods and making them more viable for clinical settings.

1.4 Methods

In this thesis, we seek to address some of the challenges that have faced previous radiomics models within lung cancer research. In the past, few nodule segmentation algorithms have demonstrated the capability to extend past LIDC-IDRI. Furthermore, there has not been a significant amount of work exploring the use of these algorithms in automated, multitask pipelines. To address these issues, we present a fully automated system for nodule segmentation and recurrence risk prediction. The initial aim of our work had been to use deep learning to advance to the novel area of recurrence risk prediction in radiomics. However, the lack of annotated imaging data with sufficient clinical information presented immediate limitations and proved to be a significant research challenge on its own. We overcome this challenge by taking advantage of the LIDC-IDRI dataset and integrating automated nodule segmentation. Thus, we circumvent the requirement for radiologist annotations and are able to utilize an unlabeled dataset for recurrence prediction.
The system we propose consists of preprocessing and two main prediction stages. In the first stage, a 2D axial slice containing nodule tissue is taken from a 3D CT series, and sent as input to a segmentation subnetwork. The predicted segmentation is then post-processed and passed to the second stage to generate a 3D region of interest (ROI) around the nodule (see Figure 1.4). The ROI is sent as input into the second subnetwork for recurrence risk prediction and assigned a risk score between 0 and 1. A cutoff of 0.5 is also applied to evaluate the system’s performance as a binary classifier.\textsuperscript{3} In the last part of our study, we build a model that synthesizes the predictions of our network with clinical data and geometric tumor features that are extracted from our predicted segmentations. We implement this final model as a random forest classifier [36]. An illustration of the entire system’s framework is provided in Figure 1.5.

\textsuperscript{3}It is important to note that while the retrospective nature of the study allows for binary classification, in clinical settings, assigning a risk score is a much more practical use to assist treatment decisions. We therefore evaluate our model on both tasks.
Figure 1.5: Proposed nodule segmentation and recurrence prediction framework. First, the primary 2D slice (slice containing the largest tumor diameter) is preprocessed to resize and normalize the image, isolate lung fields, and increase contrast. It is then fed into a segmentation network where a tumor mask is predicted and run through post processing. Next, the predicted mask is used to retrieve a 3D ROI from the original CT series, which is fed into a 3D CNN for recurrence prediction. Finally, the CNN’s output, clinical data, and geometric features of the tumor, which are extracted from the predicted segmentation, are used as features for a random forest classifier.
Figure 1.6: Residual U-Net architecture for segmentation. The output shape, in terms of image $x$ size, image $y$ size, and number of channels, is denoted below each unit. Blue arrows represent the main path of the network, dotted arrows represent residual connections, and grey arrows represent the connection between the mirrored units of the contracting and expanding paths. The colors of the layers represent different operations. A more in depth description of the network and its architecture is provided in Chapter 3.

For the first stage segmentation subnetwork, a residual U-Net architecture is used (Figure 1.6). The network is trained on the original LIDC-IDRI dataset. A more in depth description of the segmentation network and its architecture is provided in Chapter 3. For the second stage recurrence risk prediction network, a basic 3D CNN is used (Figure 1.7). It is trained on the National Lung Screening Trial (NLST) dataset from the National Cancer Institute (NCI) [19]. Since this dataset does not have any contours or other information to localize the tumor, the only validation of the segmentation results on this dataset is the performance of the overall system. A more in depth description of the recurrence prediction network and its architecture is provided in Chapter 4.
1.5 Challenges

We address several theoretical and technical challenges throughout the implementation of our framework. In both the segmentation and recurrence prediction stages, the size of the dataset is fairly limited. As a result, significant data augmentation is unavoidable. Additionally, the large problem space for segmentation poses its own difficulties. As mentioned previously, many of the state-of-the-art LIDC-IDRI segmentation algorithms rely on the provided radiologist annotations so that only a cropped region around the nodule has to be considered [9–11, 20]. However, since the NLST dataset (the dataset used for the recurrence prediction stage) does not contain any localization data aside from axial slice numbers, reducing the problem space in this way is not an option. In addition to being a more difficult task, localizing and segmenting nodules from an entire image instead of a cropped region requires significantly more computing power. It also has to handle a more dramatic class imbalance — the positive class (i.e., the pixels making up the nodule) are a significantly smaller
portion of the image than the negative class (i.e. all pixels that are not part of the nodule). In certain cases, the nodule will take up as little as 1% of the entire image. These issues are addressed by architectural modifications like adding residual connections, and using weighted loss functions. The implementation of these adjustments is discussed further in Chapters 3 and 4.

1.6 Significance

The main contributions of our work are:

1. The extension of an LIDC-IDRI trained segmentation algorithm to an independent and unannotated dataset

2. The use of the NLST dataset for recurrence risk prediction

3. A fully automated segmentation and classification system for recurrence risk prediction that is to the best of our knowledge, the first of its kind

The successful implementation of an LIDC-IDRI trained segmentation algorithm within a multitask, multi-dataset pipeline exhibits the power of current medical image segmentation. First off, it shows that these algorithms can generalize to other independent datasets. Secondly, it demonstrates that they can be integrated into classification pipelines that are trained using these different datasets. These findings indicate that a radiologist is not necessarily required to make unannotated datasets viable for classification. Rather, a potential alternative is to integrate them into a pipeline with automated segmentation. As a result, these methods can contribute to the standardization of the radiomics pipeline as it seeks to prepare for clinical use. Moreover, the breadth of challenges that can be addressed with deep learning and radiomics can be significantly expanded via reliable image segmentation. For example, in our case, we use this approach to segment the NLST dataset and implement a
recurrence risk prediction model. To the best of our knowledge, our proposed framework is the first fully automated segmentation and classification system for recurrence risk prediction.

Making the NLST dataset viable for recurrence prediction via automated segmentation is a novel development. Few datasets have the same level of diversity in their collection methods (33 different treatment centers). Even fewer have a data collection period of comparable duration (3 years of collection, 7 year follow up period). Furthermore, the extent of features included in NLST is relatively unparalleled and paramount to the success of our study. For example, the inclusion of procedural results allows us to eliminate patients that have residual disease left after surgery. The presence of residual disease calls into question whether or not the patient is ever “cancer free” and is also disproportionately correlated with recurrence [29]. Therefore, including these patients would likely skew the predictions of the neural network as images are taken before surgery. Additionally, disease progression data enables us to define a more precise outcome variable — disease free survival. In other prognostic models, the outcome is unusually based off post-surgical survival time, which usually fails to consider cause of death. Most significantly, the inclusion of clinical and treatment data (e.g. post-surgical therapy and disease staging), allows us to create a final model that integrates both imaging and clinical features.

Lastly, it is worth mentioning that seeing our framework through to clinical practice is outside our field of expertise and therefore, beyond the scope of this study. This thesis should not be interpreted in any form as an imposition of our methods on current medical practice. Rather, it should be viewed as a demonstration of the ways image processing, deep learning, and radiomics have the collaborative potential to contribute to certain areas of medicine. In the following chapters, we discuss these potential contributions alongside a more detailed description of our pipeline’s implementation.
1.7 Thesis Structure

The remainder of the thesis is organized as follows:

- **Chapter 2 – Materials and Preprocessing**: an overview of datasets used in the thesis and a discussion of the main preprocessing routine used.

- **Chapter 3 – Segmentation**: background on image segmentation and its medical applications followed by a discussion of the implementation of our segmentation framework.

- **Chapter 4 – Recurrence Risk Prediction**: background on recurrence risk prediction, prior approaches, and current challenges followed by a discussion of the implementation of our recurrence prediction framework.
Chapter 2

Materials and Preprocessing

We begin by presenting the materials and the preprocessing framework from our study. In the following chapter, we discuss the datasets used and the initial pipeline for processing their images. All code used for preprocessing is available in this project’s online repository (see Appendix B). To start, we provide an overview of each of the datasets, their collection methods, and their respective roles within our framework.

2.1 Datasets

We utilize three independent datasets throughout our study. Two are used to train and test the segmentation stage of our system. The third is used to train the recurrence risk prediction stage. In the following section, we provide an more in depth discussion each of these datasets.

LIDC-IDRI

Segmentation Training & Testing

The Lung Image Database Consortium Image Collection (LIDC-IDRI) is used to train the segmentation stage of our system. The original dataset before cleaning consists of 1,018 low-dose CT image scans from 1,010 patients in DICOM format, each with
a varying number of cross-sectional 2D slices. Additionally, each of the series has an associated XML file containing the annotations of 4 separate radiologists. The data collection was directed by the NCI and supported by the Foundation for the National Institutes of Health (FNIH) and the Food and Drug Administration (FDA). The dataset is publicly available at:

https://wiki.cancerimagingarchive.net/display/Public/LIDC-IDRI

**SPIE-AAPM**

*Segmentation Testing*

In 2015, the Society of Photo-Optical Instrumentation Engineers (SPIE) released a dataset (SPIE-AAPM) for their annual medical imaging conference challenge. We use this dataset for further tests on the segmentation stage of our system. The original dataset before cleaning consists of 83 low-dose CT image scans from 83 different patients in DICOM format, each with a varying number of cross-sectional 2D slices. Additionally, there is a csv file containing the axial slice numbers and nodule center (x, y) positions. The data was collected in a joint effort between SPIE, the American Association of Physicists in Medicine (AAPM), and the NCI. The dataset is publicly available at:

https://wiki.cancerimagingarchive.net/display/Public/SPIE-AAPM+Lung+CT+Challenge

**NLST**

*Recurrence Risk Prediction Training & Testing*

We use a subset of the NLST CT image collection to train and test our recurrence prediction stage. The subset provided contains only patients with confirmed lung cancer. Before cleaning, it consists of 1,165 sets of low-dose CT scans from 622
patients in DICOM format, each with a varying number of cross-sectional 2D slices.  

\(^1\) For each of the patients, there are anywhere between 1-3 sets of scans (one for each they were enrolled in the trial).

The NLST data collection was carried out at a network of over 33 centers across the country. The period of enrollment and primary data collection was between 2002-2004, and follow up information was completed through December 2009. In addition to the images, there is a significant amount of clinical, treatment, and outcome data that was acquired and organized into sub-datasets. The sub-datasets used for this study include:

1. **Participant Dataset**: includes demographic data, disease stage data, follow-up data, contact status, diagnostic results, mortality status, and cause of death, progression data, etc.

2. **CT Abnormalities Dataset**: contains information on each observed patient abnormality (i.e lung nodule). Features include primary axial slice, study year, etc.

3. **Treatment Dataset**: includes information about procedures during the initial course of treatment for lung cancer. Features include type of treatment administered, amount of residual disease left after surgery (or lack there of), date of procedure, etc.

More information about the NLST dataset is available at the following:

- Overview: [https://cdas.cancer.gov/nlst/](https://cdas.cancer.gov/nlst/)

- Data dictionaries: [https://cdas.cancer.gov/datasets/nlst/](https://cdas.cancer.gov/datasets/nlst/)


\(^1\)Some patients have duplicate scans with different reconstruction filters, thicknesses, etc. from the same year. The scans are therefore in sets.
2.2 Data Cleaning and Feature Engineering

LIDC-IDRI Cleaning

Given its widespread popularity, LIDC-IDRI has been used for a number of public competitions. Amongst these is the 2016 Lung Nodule Analysis Challenge (LUNA16). The LUNA16 dataset identifies and removes outliers and inconsistencies in the original data. It excludes any scans that have a particularly large slice thickness (> 2.5mm), missing slices, or inconsistent spacing. However, it also does not include full radiologist annotations. Instead of complete nodule contours, only a center point is provided. Therefore, instead of using LUNA16 as our baseline training set for segmentation, we still use LIDC-IDRI and simply eliminate all scans that are not included in LUNA16’s unique identifiers (130 eliminated in total).

Additional data cleaning is also done on a per nodule basis. First, all nodules with diameters < 3mm, which are marked in the dataset by a single point rather than a entire contour, are excluded. Second, all nodules that are not marked by the majority of radiologists (at least 3) are also excluded. In total, 888 scans and 1,187 separate nodules remain after the cleaning process.

NLST Cleaning

A significant amount of data cleaning is also done on the NLST dataset. Excluded patients include those who had non-surgical primary treatments, like chemotherapy or radiation therapy, those who had residual disease left after their surgery, and those who did not remain in contact until conclusion of the study. In order to maintain consistency with the LIDC-IDRI nodule inclusion criteria, any nodules with diameter < 4mm are also not considered.²

²Nodule diameter is not recorded in NLST if it is less than 4mm.
images require us to exclude another subset of patients. Specifically, some of the patients have multiple, but not identical, scans included for the same year. However, the abnormality dataset does not indicate which of these scans corresponds to the reported tumors. While some of the scans are virtually identical (differing only on the reconstruction filter used), others differ significantly – different slice thicknesses, different numbers of axial slices, different dates, etc. In each of these instances, the scans are inspected manually. If they are ostensibly different, the patient is excluded from the study.

**NLST Feature Engineering**

The substantial amount of clinical data present in NLST allows us to infer additional features. Most prominently, cancer recurrence (our target outcome variable) is inferred from the progression data and treatment data. If the patient had a cancer progression, the time of the progression is compared to the time of surgery (where time is recorded in days since randomization). If the patient is found to have a cancer progression after surgery, they are marked as having recurrent cancer. If the progression occurred before the surgery or there is no progression, they are marked as recurrence free. Curiously, in doing this, we also find that for several patients, there is no recorded post-surgical progressions, but there is still a recorded death due to lung cancer. It is unclear if this is due to misreported/missing data or another lung cancer related complication, so they are excluded from the study. Lastly, we infer an adjuvant therapy feature. If the patient has any documented chemotherapy treatments in addition to surgery, we compare the time the treatment is administered to the time of surgery. If it is determined that chemotherapy was administered post-surgically, the patient is marked as having received adjuvant therapy. The importance of this

---

3 Randomization is the point at which enrolled trial subjects are assigned to treatment/control groups.
adjuvant therapy feature is discussed in Chapter 4.

In total, after cleaning and feature engineering, there are 331 patients remaining. From these patients, there are 899 recorded nodules. There are 242 patients and 686 nodules that are part of the no-recurrence class, and 89 patients and 213 nodules that are part of the recurrence class.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Purpose</th>
<th>Important Attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIDC-IDRI</td>
<td>segmentation training and testing</td>
<td>radiologists’ annotations, primary axial slices</td>
</tr>
<tr>
<td>SPIE-AAPM</td>
<td>segmentation testing</td>
<td>nodule center of masses, primary axial slices</td>
</tr>
<tr>
<td>NLST</td>
<td>recurrence risk prediction training and testing</td>
<td>disease progression, residual disease after surgery, treatments administered, cause of death, cancer stage, demographic information, adjuvant therapy, primary axial slices</td>
</tr>
</tbody>
</table>

Table 2.1: Dataset summary

2.3 Preprocessing

The images from all datasets are passed through the same preliminary preprocessing pipeline. Each stage of the pipeline is described below. An illustration is also provided in Figure 2.1.

1. First, we extract the cross-sectional 2D image slices containing tumors from the 3D series.

2. The extracted slices are then normalized and resized from 512 × 512 pixels to 256 × 256 pixels.

3. The normalized and resized image is then passed through a lung field (not nodule) segmentation process to remove the background. This process consists
Figure 2.1: Preprocessing pipeline. First, the primary 2D slice (slice containing the largest tumor diameter) is extracted. Second, the slice is resized and normalized. Third, the lung fields are isolated. Last, contrast is enhanced.

of two steps. First we use K-Means clustering and binary thresholding to select the main lung regions. Second, a series of erosions and dilations is applied to include the anatomical structures within these lung regions (including any nodules) without adding back excessive amounts of background.

4. Finally, the image goes through contrast enhancement.

For a small subset of the images, effective preprocessing proves to be challenging. For example, when the tumor is attached and closely embedded in the lung wall, the lung field segmentation step excludes a significant portion of the tumor from its output. To address this, images are inspected manually and those facing this issue are run through additional dilations (see Figure 2.2). However, as is also evident from Figure 2.2, these extra dilations lead to an increase in background noise. For a few images, the amount of extra background is particularly excessive, so they are discarded. Additionally, a small number of images become overexposed during contrast enhancement. The resulting image contains pixels that are essentially binary. That
Figure 2.2: Image with challenging lung field segmentation. (a) Original raw image. (b) Target nodule mask. (c) Applying a lung field mask to remove background and surrounding tissue, but also removing the tumor. (d) Applying additional dilations to include the tumor.

is, everything inside the lung field is white, and everything outside is black. These images are also discarded.

2.4 Segmentation Target Generation

In addition to input processing, a binary target mask needs to be generated for the segmentation subnetwork (see figure 2.3). To do so, the lists of pixels corresponding to nodule boundaries are extracted from the LIDC-IDRI XML files. The internal areas of these boundaries are then computed and used to produce binary masks. Since each nodule has multiple boundaries (one from each radiologist), we take the union of the masks produced using each boundary. When tested within the full segmentation pipeline, we find that a preprocessing routine that takes the union of the masks (as
Figure 2.3: Binary target generation for segmentation. (a) The original raw image. (b) The nodule boundaries plotted using the marked pixels provided in the radiologist annotations. (c) Binary target mask.

opposed to the intersection or average) is the most effective. The remainder of the segmentation pipeline is discussed in the following chapter.
Chapter 3

Lung Nodule Segmentation

In this chapter we present further background on image segmentation, its medical applications, and discuss the implementation of the segmentation stage of our system. Towards the end of the chapter, we also discuss the experiments run to test our model’s performance and compare our results the current state of the art.\footnote{All code used to implement the segmentation stage and run experiments is available in this project’s online repository. See Appendix B for more details.} We begin by giving an overview of the task posed by image segmentation.

Stage 1: Segmentation

Figure 3.1: Segmentation Pipeline. After preprocessing, images are sent to the segmentation stage of our system. First, they are input to a CNN (a residual U-Net) to predict the tumor segmentation. The predicted segmentation is then run through post processing to get the final prediction.
3.1 Background on Lung Nodule Segmentation

3.1.1 Overview of Image Segmentation and its Medical Applications

Semantic segmentation, or simply segmentation, is a common problem seen in image processing. Unlike classification, the task of labeling an image, or even classification with localization, which is the task of labeling and locating the object being labeled, segmentation seeks to label each pixel within the image (see Figure 3.2). Recent breakthroughs continue to tackle this problem for a variety of applications [7–11, 37].

Figure 3.2: Example of Image Segmentation. The figure above provides an example of the difference between classification, classification with localization, and segmentation. While classification seeks to label the image, and classification with localization seeks to label and locate the subject of the image (generally with a bounding box), segmentation seeks to label each pixel within the image.
Medicine in particular is an area that continues to benefit from segmentation enabled technology. The medical community’s heavy use of diagnostic imaging like magnetic resonance imaging (MRI), CT, and mammography, has led to a dependency on these algorithms to facilitate contouring anatomical structures. The process is critical in a variety of applications including, but not limited to, the identification of cellular structures in electron microscopy images, pathology localization, and the estimation of tissue volumes.

Amongst the more specific medical applications is lung nodule segmentation. In recent years, significant research effort has gone into developing automated methods for segmenting lung tumors in order to aide the radiology process. While certain semi-automated methods have enabled software that facilitates radiologist led segmentation, there is still a need for full automation. This has become especially true due to the current state of lung cancer screening. In the past decade, significant evidence has shown that using low-dose CT scanning as opposed to chest radiography (i.e. X-ray) for screening reduces lung cancer mortality rates [38]. These findings have led to recommendations from the American College of Radiography [39], the US Preventive Services Task Force [40], and the National Comprehensive Cancer Network [41] reemphasizing that CT screens should be the first step in managing lung abnormalities. As a result, CT screening has become the default modality in lung cancer imaging. However, the manual analysis of a CT scan, which will usually contain hundreds of slices, is an extremely laborious process. Not only does this lead to less efficient clinical protocols, but it also makes building large scale, manually segmented datasets extremely difficult. However, as is demonstrated by this study, these issues can potentially be resolved by the development and integration of automated nodule segmentation.
3.1.2 Review of Image Segmentation Methods

There are several techniques that have historically appeared in the literature on image segmentation. These same techniques are seen across different applications, including those within the medical field. While many modern studies have synthesized and refined these approaches, we provide a general overview of the fundamental methodology behind each of them. The list provided is not exhaustive but includes most of the common overarching categories [42–44].

Thresholding

Thresholding is usually considered the most basic method for segmentation. In medical imaging, it has a history of being used for a variety of tasks. Examples include lung field segmentation (similar to the one used in our preprocessing routine) [45] and the segmentation of different regions in the brain [46]. The fundamental aim of thresholding methods is to define some pixel intensity (the threshold) around which to partition the classes. The segmentation is determined by labeling all pixels above the threshold as one class, and all pixels below as the other. The value of the threshold is usually determined via manual inspection of histograms. More advanced variants of thresholding are also often used. Commonly seen examples are multiple thresholding, where more than one threshold is used to partition the classes, and variable thresholding, which determines the threshold as a function of the pixel’s coordinates. Figure 3.3 illustrates the process of basic thresholding and multiple thresholding.

Edge Based Approaches

Despite the simplicity and interpretability of thresholding methods, raw pixel intensity alone does not usually convey enough information for advanced segmentation tasks. One alternative is edge based segmentation, which takes advantage of the fact
that there is often a rapid drop off or increase in pixel value at the region boundaries. They have been used for the segmentation of tumors across cancer types and imaging modalities [47, 48]. Instead of selecting a predefined threshold for pixel values, these methods find region edges based off the gradients of pixel intensity. While alone, these techniques only find boundary pixels and not a full segmentation, they are often used in conjunction with region growing techniques [48].

Region Based Approaches

Region based methods are one of the most common techniques used for segmentation in the medical field [48–51]. These types of approaches attempt to partition image pixels by finding connected components. Pixel “connectedness” is defined by some predefined criteria like similarities in pixel values or edge information. A significant caveat however, is that region growing methods almost always require the manual definition of a seed point. To overcome this, there are many variants of region based approaches, like combination with edge based methods [48], split and merge algo-
Figure 3.4: Example of region growing for segmentation. Above, region growing is used for the semi-automated segmentation of white matter brain tissue. A seed point is first planted manually, and the rest of the region is automatically segmented by finding “connected” pixels [52].

Rithms [49], and watershed segmentation [49–51]. Split and merge algorithms splits the image into quadrants, merging adjacent, similar quadrants, and recursively splitting dissimilar quadrants. Watershed segmentation is based off a topological interpretation of an image where high pixel values are conceptualized as ridges, and low pixel values are conceptualized as basins. When the image is “flooded” to a certain level (hence the name watershed), “basin” regions are merged, and predefined edges can be used as “dams” to prevent the accidental merging of dissimilar regions.

Clustering

Similar to region based approaches, clustering is another common technique for segmenting images based off pixel similarity. It is used in the medical field for a variety of tasks [37, 50, 53, 54]. These approaches segment images by iteratively grouping pixels into clusters, characterizing the properties of each cluster, and reassigning the
pixels based off their similarity to each of the clusters’ properties. Common methods include the k-means algorithm (also used within our preprocessing pipeline) [50, 53], the fuzzy c-means (FCM) algorithm [55], and the expectation-maximization (EM) algorithm [54]. While these techniques are effective for a number of tasks, a common barrier for both clustering and region based techniques is sensitivity to parameter initialization. For certain applications, optimal parameter values can vary significantly between images. Moreover, both of these techniques depend heavily on well defined regions and do not generalize well when region boundaries are not distinct.

Deep Learning

Recently, many of the most effective approaches for medical image segmentation have come from deep learning, the subset of machine learning focuses on neural networks [7–12]. In particular, CNNs continue to achieve state of the art results for a wide variety of tasks. Medicine is no exception, as variants of CNNs continue to solve segmentation problems across medical domains. As we will demonstrate with lung nodule segmentation, CNNs are able to overcome many of the barriers facing other common segmentation methods.

3.1.3 Complexity of Lung Nodule Segmentation

Despite the abundance of tools for image segmentation, the internal structure of chest CTs makes lung nodule segmentation a highly non-trivial task. The distribution of pixel values and lack of homogeneity between nodules makes it difficult to apply many of the common methods. Generally, the typical axial slice of a CT image capturing a lung nodule consists of four main regions (see Figure 3.5a):
Figure 3.5: Lung CT regions and histograms. (a) Original images and their respective regions. (b) Histograms of the original image (excluding background). (c) Density of pixel values in each of the non-background regions. While (b) demonstrates the bi-modal distribution typical in lung CTs, it is clear from (c) that pixels of nodule tissue are usually “blended” between the peaks of this distribution and can vary significantly between images.

1. Region 1: background (air).
2. Region 2: the non-lung region of the body – this region usually includes heart tissue, vertebrae, torso muscles etc.
3. Region 3: the lung region – this region includes lung tissue, blood vessels, and other anatomical structures within the lung.
4. Region 4: the nodule region – a subset of the lung region.

While these regions may sometimes be distinguishable from one another by eye, there is no clear cutoff between the pixel values of each region. More often than not, the entire image (excluding background) follows a bi-modal distribution with peaks that roughly distinguish region 2 from region 3 (see Figure 3.5b). However, the pixel
Figure 3.6: Lung nodule attenuation. (a) Solid nodule with well defined margins. (b) Ground-glass nodule with poorly defined margins. Often, the lack of margin definition in ground-glass nodules poses a challenge for segmentation algorithms.

values of the nodule, which is the region of interest, are usually “blended” somewhere in between the two peaks and can vary significantly between images (see Figure 3.5c). This makes thresholding approaches, or any other approach that is heavily dependent on pixel intensity, poorly equipped to handle this type of segmentation.

Beyond raw pixel values, the attenuation and location of certain nodules can add additional complications. While some nodules are solid and well defined, others have ground-glass (non-solid) opacities or mixed opacities (see Figure 3.6). The poorly defined margins of these ground-glass nodules pose a challenge for many segmentation methods. For edge detection algorithms in particular, the lack of a steep gradient in pixel value often causes them to go undetected. Different types of challenges can also occur for solid nodules. It is not uncommon for these nodules to be embedded in the lung wall or other surrounding structures (see Figure 3.7). Therefore, it becomes virtually impossible to define a clear boundary between nodule and non-nodule tissue. These issues often persist even if a prepossessing routine like the one in our study is able to isolate the lung field from most of the surrounding tissue. The series of
erosions and dilations in the lung field isolation process will also often include at least some residual, surrounding tissue along with the embedded tumor. As a result, edge based, region based, and clustering methods all struggle in these cases [56].

Lastly, many approaches (i.e. thresholding, edge based, region based, and clustering) often fail to address that there is a strong similarity between nodule tissue and other anatomical structures found within the lung (see Figure 3.8). For fully automated methods where no prior seed point is defined, it is crucial to distinguish between the two. While some models try to address this by using a classifier to remove false positives after the initial segmentation, they still tend to be outperformed by deep learning methods. For these reasons, we choose to use a neural network for our segmentation model.
Figure 3.8: Similarities between lung nodules and other anatomical structures within lungs. The ability to discriminate between nodule and non-nodule structures is critical for segmentation, but difficult for many non deep learning methods.

3.2 Related Work

As we discuss in the previous section, lung nodule segmentation is a highly non-trivial task. Nodules can be poorly defined and can resemble other anatomical structures within the lung. There is also a lack of homogeneity between different nodules. To address these challenges, we decide to use a CNN for the segmentation stage of our system. The specific design of this CNN is inspired by two previous image segmentation studies in particular [7, 37]. These studies are discussed in the following section.
3.2.1 U-Net

For our network architecture, we use a variant of the U-Net [7]. Originally developed for the segmentation of neuronal structures in electron microscopy stacks, the U-Net outperforms prior methods significantly and has become the state of the art for many segmentation problems. The model’s architecture consists of a contracting path (left in Figure 3.9) and an expanding path (right in Figure 3.9). The contracting path resembles a typical CNN where the image’s spatial information is reduced through unpadded \((3 \times 3)\) convolutions and max pooling operations. The expanding path mirrors the contracting path but replaces the pooling layers with upsampling to restore the original image dimensions. Additionally, at each expansion stage, a concatenation with the corresponding contracting unit is applied. These concatenations facilitate the propagation of high level features and tend to improve localization.

**Figure 3.9:** The original U-Net architecture as seen in [7]. Figure as described in [7]: “Each blue box corresponds to a multi-channel feature map. The number of channels is denoted on top of the box. The x-y-size is provided at the lower left edge of the box. White boxes represent copied feature maps. The arrows denote the different operations.”
3.2.2 Residual U-Net

An elegant extension of the original U-Net is presented by Zhang et al. [37] for the segmentation of roads in aerial images. Entitled the ResUnet, the proposed model adds residual connections to initial architecture, an idea that was first presented by He [57]. In essence, the residual unit is constructed by adding the unit’s original input to its output (see Figure 3.10). Thus, the inputs and outputs of each of the units can now be represented in the following form:

\[
x_k = f(y_{k-1}) \\
y_k = F(x_k) + h(x_k)
\]  

(3.1)

where \( f \) is the activation function, \( x_k \) is the input to the current unit, \( F \) is the function representing the unit’s internal layers (in our case convolutions), \( h \) is an identity mapping, and \( y_k \) is the output of the current unit (see Figure 3.10). These residual connections are a proven effective way to alleviate vanishing gradients and degradation issues that occur when training deep networks.

**Figure 3.10: Residual Unit Architecture** Originally presented in [57], residual units are constructed by adding the unit’s original input to its output where \( x_k \) is the input to the unit, \( F \) is the function representing the unit’s internal layers (in our case convolutions), \( h \) is an identity mapping, and \( y_k \) is the unit’s output.
3.3 Proposed Method

In the following section, we present our full segmentation pipeline. We discuss our proposed network, the training process, and the preprocessing routine used to refine predictions. We begin with a description of the network’s architecture.

3.3.1 Proposed Network Architecture

For our lung nodule segmentation model, we use an adapted version of the ResUnet [37]. Our choice is motivated by difficulties encountered with the original U-Net. We observed that at a certain point in training the original U-Net architecture, performance steeply drops. In particular, recall often converges to zero. (See Equation 3.9 in Section 3.4.1 for a definition of the recall metric). This issue is likely caused by the significant class imbalance of nodule to non-nodule pixels. In certain images, the nodule makes up less than 1% of the image.

The architecture we present contains two key modifications. First, we adapt the bridge between the contracting and expanding paths of the network. Unlike the other units in the proposed model, the bridge unit does not contain any residual connections. Instead, it closely resembles its counterpart in the original U-Net. Secondly, a mask from the original input is applied to the final output of the network that eliminates any false positives detected outside of the lung region.

The network architecture is illustrated in Figure 3.11. It can be best understood by inspecting the individual units within the contracting, bridge and expanding sections respectively.

1. **Contracting Units**: Similar to [37], each residual unit within the contracting path applies a batch normalization and ReLU activation to the previous unit’s output. This is followed by a $(3 \times 3)$ convolution with a stride length of 2. The
Figure 3.11: Architecture of the proposed residual U-Net for segmentation. The output shape, in terms of image \(x\) size, image \(y\) size, and number of channels, is denoted below each unit. Blue arrows represent the main path of the network, dotted arrows represent residual connections, and grey arrows represent the connection between the mirrored units of the contracting and expanding paths. The colors of the layers represent different operations.

The process is then repeated a second time, but with a convolution stride length of 1. Finally, the original input is added to the output of the second convolutional layer and passed on to the next unit. Unlike the original U-Net, the stride length of 2 in the first convolution serves as the downsampling operation, so no max pooling operation is applied.

2. **Bridge Unit**: Next, the bridge unit consists of this same batch normalization, activation, convolution sequence. However, the sequence is applied only once before the output is passed to the expanding path. Also, there is no residual connection across the unit.
3. **Expanding Units**: The units in the expanding path have a similar structure to the contracting units. However, first, a $(2 \times 2)$ up-sampling operation and concatenation with the output of the mirrored contracting unit is applied. Also, both convolutional layers in the expanding path have stride lengths of 1.

Following the expanding path, the image is passed through a $(1 \times 1)$ convolution and sigmoid activation. This activation is then followed by the application of a binary mask to eliminate any false positives detected outside of the lung region. The mask is generated from the network’s original input. Specifically, since our preprocessing routine yields a normalized image with the background removed, applying a ceiling operation to this image creates a binary mask of the lung field. Then, applying the mask to the output of the sigmoid removes any falsely activated pixels from outside of the lung region. More simply put:

$$\hat{Y} = [X] \cdot \sigma(X)$$

where $\hat{Y}$ is the network’s final predicted segmentation, $X$ is the initial input to the network, and $\sigma(X)$ is the output from the network’s final sigmoid activation. This binary mask proves to be an effective way to eliminate noise and helps the network converge faster.

### 3.3.2 Training

Following the preprocessing routine discussed in Chapter 2.3, the images are partitioned randomly into training, testing, and validation sets. In order to prevent different images from the same patient appearing in more than one of these sets, we first create a random split based off patient ids. We then partition the corresponding images accordingly. In order to maximize the number of images for training, all slices
containing nodule tissue are included in the training and validation sets. This means that multiple images from each patient, and even multiple images from each nodule, are usually included. For the test set however, only the image with the largest cross-sectional tumor diameter is considered. The choice to not include more images per patient in the test set is made for two reasons. First, we want to prevent skewing the final evaluation of our model towards patients with more nodules/nodule slices. Second, during recurrence prediction in the next stage of the system, only slices with the largest cross-sectional tumor diameter are used to segment each nodule. The final breakdown is 5,013 images in the training set, 500 in the validation set, and 138 in the test set.

The model is implemented using the Keras 2.2.4 using tensorflow 1.14.0 as the backend. Due to the significant imbalance of non-nodule to nodule pixels, weighted binary cross-entropy is used as the loss function:

\[
L_{wbce} = -\frac{1}{N} \sum_{i=1}^{N} w \cdot y_i \cdot \log(h_{\theta}(x_i)) + (1 - y_i) \cdot \log(1 - h_{\theta}(x_i))
\] (3.3)

where \(N\) is the number of samples, \(w\) is the weight given to the positive class, \(y_i\) is the true label for the \(i\)’th sample, \(h_{\theta}\) is the network with weights \(\theta\), and \(x_i\) is the input for the \(i\)’th sample. We use a weight of 12.0 for the positive class. For our optimizer, we use the Adam optimizer with a global learning rate of 0.0001. The training process runs for a maximum of 20 epochs with 1,000 steps per epoch and a batch size of 12. The model is saved after each epoch and validation loss is monitored. The training process is terminated if a notable decline in validation performance occurs.

Another critical element of our training process is data augmentation. Due to the limited number of samples and the similarity between images from the same patient, augmentation is crucial to avoid overfitting. All augmentation is done in real time during training instead of as part of preprocessing. The operations include random
Figure 3.12: Segmentation network data augmentation. After preprocessing, images and labels are augmented via this process in real time during training. Each image shown above illustrates only one of the possible results from the augmentation process. The magnitude of each transformation is not fixed, but rather chosen randomly during training. Likewise, whether or not to apply the horizontal flip is also chosen randomly. After all augmentation transformations are applied, the image is normalized to the 0-1 range and input to the network.

Brightness shifts, rotations, shears, zooms, and horizontal flips. (See Figure 3.12). The magnitude of each of the operations is chosen randomly during training, and all images are also normalized to the 0-1 range at the end of the augmentation process.

3.3.3 Post-Processing

After the final output from the network, a small amount of post-processing is done. For certain predicted segmentations, there are multiple activated regions. This can be due to the presence of more than one nodule in the same image, or false positives. In order to both simplify the transition to the next stage of the system (recurrence classification on a single nodule) and eliminate potential false positives, we select only one of these regions to remain activated in the final segmentation. The region selected is the most densely activated region. Thus, the following process is used to update the predicted segmentation:

1. Find the connected components of the network’s predicted segmentation (i.e.
get the pixel coordinates corresponding to each of the activated regions). Pixels are considered connected if they are direct neighbors in the $x,y$ direction and have value $> 0.1$.

2. Find the most densely activated region:

$$r^* = \arg \max_{r_k \in R} \left\{ \frac{\sum_{i,j \in r_k} \hat{Y}_{i,j}}{|r_k|} \right\}$$  \hspace{1cm} (3.4)

where $\hat{Y}$ is the (non-binary) output from the network, each $r_k$ is a list of pixel coordinates for one of the activated regions, $R = \{r_1, r_2, ..., r_n\}$ is the set of all of these lists, and $r^*$ is the most densely activated region.

3. Set all activated pixels not within the most densely activated region to 0:

$$\hat{Y}'_{i,j} = \begin{cases} Y_{i,j} & \text{if } (i,j) \in r^* \\ 0 & \text{o.w.} \end{cases}$$  \hspace{1cm} (3.5)

where $\hat{Y}'$ is the updated predicted segmentation after post-processing.

Although this post-processing scheme can potentially eliminate true positives (especially when there are multiple tumors present as is evident in Figure 3.13), we find that it still improves the system’s overall performance. Moreover, for recurrence prediction in next stage, the ROI of an individual nodule is input from each image. Thus, retaining only one activated region per image actually facilitates the transition between stages.
Figure 3.13: **Post-processing.** After the predicted segmentation is output from the network, we modify it so that only the most densely activated region is included. This both improves overall model performance by eliminating false positives and eases the transition to the recurrence prediction stage of the system.

### 3.4 Results and Experiments

#### 3.4.1 Evaluation Metrics

We evaluate our model using several metrics that compare the predicted segmentations with the manual segmentations made by the radiologists:

1. **Dice coefficient:**
   
   Dice coefficient measures the relative overlap between two images. It is defined as:
   
   $$\text{Dice}(X, Y) = \frac{2(|X| \cap |Y|)}{|X| + |Y|}$$  
   
   where $\cap$ is the intersection operator and $X$ and $Y$ are the two images in question. A dice coefficient of 1 occurs when there is a perfect overlap. A dice coefficient of 0 occurs when there is no overlap.

2. **Percent of primary nodule included in ROI:**
   
   We define percent of primary nodule included in ROI as the maximum percentage of *any* tumor included in the $50 \times 50$ ROI around the center of the
Figure 3.14: Computing percent of primary nodule included in ROI. The percent of the primary nodule included in ROI is one of the primary metrics used to evaluate the model. It measures the (maximum) percentage of a nodule that is included within the $50 \times 50$ ROI generated from the predicted segmentation. Although unconventional for evaluating segmentation performance, we use this metric because the ROI is an integral part of our system’s next stage. It can be expressed as:

\[
ROI = \{\bar{r} \times -25, \bar{r} \times +25\}, \{\bar{r} \times -25, \bar{r} \times +25\} \\
\text{PercentIncluded} = \max_{r \in R} \left\{ \frac{r \times ROI}{|r|} \right\} 
\]

(3.7)

where each $r \in R$ is a set of pixel coordinates for a nodule in the true mask, $\bar{r}$ is the pixel coordinates of the predicted activated region (following post-processing), and $\bar{r} \times$ and $\bar{r} \times$ are the average $x$ and $y$ coordinates of this predicted region (i.e. its center). For further clarification, an illustration is provided in Figure 3.14.

3. Precision and Recall:

The last two metrics we use are pixel-wise precision and pixel-wise recall. Precision is defined as the ratio of correctly predicted positive instances to all predicted
positives instances:

\[ \text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \] (3.8)

Recall is defined as the ratio of correctly predicted positive instances to all actually positive instances:

\[ \text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \] (3.9)

We find these two metrics to be significantly more practical than accuracy, which becomes extremely inflated as a result of the class imbalance.

3.4.2 Results

To test the performance of our model, we compare it with two other architectures trained within the same pipeline. The architectures chosen are the original U-Net [7] and the original ResUnet [37]. All architectures are tested using the same preprocessing, training, and post-processing routines, as well as the same training, validation, and test sets. The only change made is to the weight of the positive class in the weighted binary cross-entropy loss function in the original U-Net. A weight of 65.0 has to be applied in order to prevent the network from converging to all negative predictions. Table 3.1 shows a comparison of the results on the test set before post-processing, and Table 3.2 shows a comparison of the results on the test set after post-processing. Since a single activated region must be isolated to compute the ROI, percent of primary nodule cannot be computed before post-processing. We include it separately in Table 3.3.

The results in Tables 3.1, 3.2, and 3.3 validate the benefits of our proposed architecture. The significant increase in performance seen between the original U-Net and
### Table 3.1: Segmentation Results Before Post-Processing.
Comparison of the segmentation performance of three different architectures on the test set before post-processing.

<table>
<thead>
<tr>
<th>Architecture</th>
<th>Dice Coefficient</th>
<th>Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed Model</td>
<td>67.7%</td>
<td>81%</td>
<td>64%</td>
</tr>
<tr>
<td>Original ResUnet</td>
<td>64.2%</td>
<td>76%</td>
<td>66%</td>
</tr>
<tr>
<td>Original U-Net</td>
<td>50.9%</td>
<td>92%</td>
<td>38%</td>
</tr>
</tbody>
</table>

### Table 3.2: Segmentation results after post-processing.
Comparison of the segmentation performance of three different architectures on the test set after post-processing.

<table>
<thead>
<tr>
<th>Architecture</th>
<th>Dice Coefficient</th>
<th>Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed Model</td>
<td>70.3%</td>
<td>78%</td>
<td>69%</td>
</tr>
<tr>
<td>Original ResUnet</td>
<td>64.4%</td>
<td>73%</td>
<td>68%</td>
</tr>
<tr>
<td>Original U-Net</td>
<td>54.3%</td>
<td>76%</td>
<td>38%</td>
</tr>
</tbody>
</table>

### Table 3.3: Segmentation results for percent of primary nodule included in ROI.
Comparison of the average percent of primary nodule included in the ROI by each of the architectures on the test set.

<table>
<thead>
<tr>
<th>Architecture</th>
<th>Percent of Primary Nodule Included</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed Model</td>
<td>90%</td>
</tr>
<tr>
<td>Original ResUnet</td>
<td>87%</td>
</tr>
<tr>
<td>Original U-Net</td>
<td>80%</td>
</tr>
</tbody>
</table>

the original ResUnet demonstrates the efficacy of adding residual connections. By tailoring the residual U-Net architecture to the task of lung nodule segmentation, we are able to achieve a performance competitive with other state of the art approaches in this domain.

### 3.4.3 Further Comparisons

We also benchmark our approach against the recent lung nodule segmentation literature. A wide variety of models are used for this comparison. Specifically, approaches include a combined active contour and Fuzzy C-means model [58], a region based
<table>
<thead>
<tr>
<th>Study</th>
<th>Approach</th>
<th>Dataset</th>
<th>Dice Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed Model</td>
<td>Adapted Residual U-Net</td>
<td>LIDC-IDRI</td>
<td>70.3%</td>
</tr>
<tr>
<td>Kamal et al. [8]</td>
<td>Recurrent 3D U-Net</td>
<td>IEEE 2018 VIP Cup</td>
<td>74.0%</td>
</tr>
<tr>
<td>Nithila et al. [58]</td>
<td>Region Based CNN</td>
<td>LIDC-IDRI</td>
<td>60.0%</td>
</tr>
<tr>
<td>Wang et al. [12]</td>
<td>Active contour and Fuzzy C-mean</td>
<td>LIDC-IDRI</td>
<td>60.0%</td>
</tr>
</tbody>
</table>

Table 3.4: Comparison with other segmentation models.

CNN [12], and a recurrent 3D U-Net [8]. The results are shown in table 3.4.\(^2\)

With the exception of the Recurrent 3D U-Net proposed by Kamal et al. [8], our model outperforms the other methods. It is unclear if the use of different datasets has a significant effect on these results ([8] is the only study not using LIDC-IDRI), but our approach still proves to be competitive. Moreover, unlike previous approaches, we validate the success of our model further by testing it on an independent dataset. These tests are discussed in the following section.

### 3.4.4 SPIE-AAPM Testing

While the lack of fully contoured datasets aside from LIDC-IDRI makes testing our segmentation model on an independent dataset difficult, we still test its relative performance on the SPIE-AAPM dataset. While it does not contain full nodule contours, SPIE-AAPM does provide a nodule center point for each image. Therefore, we compute the distance from the center of our predicted ROI to this labeled nodule center point (see Figure 3.15). We find that for 85% of the images, our proposed model is able to predict an ROI center within 10 pixels of the labeled center. However, it is worth noting that our model might actually perform better than this number.

\(^2\)Other studies have reported dice scores as high as 82% but reduce the problem space to a cropped region around the nodule [11]. Therefore, we do not consider them as a benchmark.
Figure 3.15: SPIE-AAPM testing. An example of the results of the proposed segmentation model using the SPIE-AAPM dataset. The raw image, provided nodule center point, and corresponding encircled nodule is shown on the left. The image after preprocessing that is input into the proposed segmentation network is shown in the middle. The predicted segmentation is shown on the right.

suggests. SPIE-AAPM was initially released for a malignancy classification challenge (as opposed to a nodule detection challenge) and only provides one nodule center per image. Therefore, if our prediction is off when there are actually multiple nodules present, there is no way of knowing if our model detects an unlabeled nodule or simply a false positive. Regardless, putting the granularity of these results aside, these tests still demonstrate the capability of our model to generalize to an independent dataset – a significant finding before advancing to the recurrence prediction stage.

3.5 Discussion

By taking advantage of methods from both inside [7] and outside [37] the medical community, we are able to implement a model for lung nodule segmentation that is competitive with the current state of the art. Additionally, by testing its performance on the SPIE-AAPM dataset, we are able to confirm that it is robust. Ultimately however, for our purposes, the most significant way to validate our model is to demonstrate its efficacy in our full pipeline. This is discussed in the following
chapter, where we present a recurrence risk prediction system that generates its input from our model’s predicted segmentations.
Chapter 4

Recurrence Prediction

In the following chapter, we discuss the recurrence prediction stage of our system. We begin by reviewing the challenge posed by lung cancer recurrence and the current approaches that are used to address it. We then continue into a discussion of our proposed recurrence prediction system and its integration of automated nodule segmentation.¹

Stage 2: Recurrence Prediction

Figure 4.1: Recurrence prediction pipeline. First, the predicted segmentation from the previous stage is used to generate a 3D ROI around the identified tumor. Next, this ROI is passed as input to a recurrence risk prediction CNN. Finally, the CNN’s output, clinical data, and geometric features of the tumor, which are extracted from the previous stage’s predicted segmentation, are used as features in a random forest classifier.

¹All code used to implement the recurrence prediction stage and run experiments is available in this project’s online repository. See Appendix B for more details.
4.1 Background on Recurrence Prediction

4.1.1 Overview of Recurrence in Lung Cancer

Disease recurrence is a significant challenge in the treatment of post-surgical lung cancer patients. Improved detection can enable an earlier diagnosis and improve the chance of complete surgical resection (the only known curative treatment for lung cancer [59]), but depending on stage, 30% to 55% of patients will have their cancer return [3] (see Figure 4.2). On average, these patients experiencing recurrent cancer do not survive more than a year [3]. These high mortality rates due to tumor recurrence is one of the main reasons why lung cancer causes significantly more deaths each year than any other form of cancer. [1].

![Figure 4.2: Rates of lung cancer recurrence.](image)

The plot above shows post-operative survival rates by cancer stage. The data is from 760 surgically treated lung cancer patients that were enrolled in the Environment and Genetics in Lung Cancer Etiology study, 2002–2005. Figure produced by [2].

51
Post-operative treatment options for lung cancer patients can vary significantly and often highly depend on their evaluated risk for recurrence. One of the most common courses of treatment is adjuvant therapy – treatment, usually chemotherapy, that is administered in addition to surgery. Although adjuvant therapy has been the standard for treating certain cancers like breast and colorectal cancer, its effectiveness for post-operative lung cancer patients has only recently been shown. \[60\]. Significant research effort continues to go into confirming its efficacy for patients in various stages \[61\]. However, adjuvant therapy is not without its own disadvantages. In addition to detracting from the patient’s quality of life, adjuvant therapy is not always safe. While uncommon, there is even a risk of treatment-related mortality especially for patients recovering from surgery \[59\]. Its effectiveness can also vary between patients and the decrease in incidence of recurrence is not always dramatic \[61\]. Therefore, depending on a patient’s evaluated risk of recurrence, other options like rigorous surveillance or non-surgical primary treatments can be considered. Prognostic tools that more accurately predict this risk can help physicians make choices that optimize outcome and quality of life on a per patient basis. In the age of precision medicine, the expressed need for this type of individualized treatment is growing significantly.

4.1.2 The TNM System

The main prognostic tool used to determine cancer treatment guidelines is the TNM staging system. Aside from the extent of residual disease left after surgery, this system is the primary method used to stratify patients into risk groups for recurrence \[29\]. The TNM system stages cancer based off size and extent of tumor (T), presence of cancer in regional lymph nodes (N), and metastasis (M). Often, the TNM combinations are grouped into 5 less detailed stages. Stage 0 denotes that abnormal cells are present but that they are not cancerous. However, they may become cancerous. Stages I–III denote that cancer is present. The size of the tumor and the extent it
has spread to nearby tissue determines the staging number within this range. Stage IV denotes that cancer has spread to distant parts of the body [62]. While patients in later stages tend to have a higher chance of developing recurrent disease, it has been shown that the TNM system is far from a sufficient framework for more granular predictions [29]. Even when there is no residual disease left after surgery, incidence of recurrence can vary significantly between patients in the same stage [59]. Moreover, patients in early stages also often develop recurrent tumors [63–65]. Specifically, post-surgical recurrence rates for stage I patients have been reported to range from 27% to 38% [65]. However, with the TNM system, stratifying a patient as high or low risk for recurrence (and the subsequent choice of whether or not to administer of adjuvant therapy) is usually determined by whether or not their cancer has progressed to stage II [29]. Thus, there is a persistent need for a more granular prognostic tool, and in particular, for one that can more accurately identify high risk stage I patients or low risk stage II patients [29].

4.2 Related Work

For decades, there has been significant research dedicated to finding a more advanced prognostic tool for recurrence risk prediction. The increased interest in personalized medicine has motivated these efforts even further [30, 66]. Studies have primarily been spread across several fields:

1. **Clinical Risk Factors:**

   Many studies have tried to build models identifying clinical factors beyond just stage that are related to recurrence [21, 22]. Factors commonly used include age, sex, race, treatment, stage, extent of disease, and smoking status.

2. **Histology and Immunohistochemistry:**
Several studies have also shown that advanced histological methods can be utilized for analyzing risk of recurrence via the examination of protein expressions [24, 25]. Specifically, the use of immunohistochemistry (IHC), which is integration of histological, immunological, and biochemical techniques, has yielded promising results [25].

3. Gene Expression Analysis:

Additional research has focused on identifying gene expressions that are correlated with recurrence [26–28]. Efforts have focused on both the molecular classification of lung tumors themselves [26, 27], and global gene expression profiling [28].

Despite the promise of some of these studies, there have been significant obstacles inhibiting their entrance to the clinical setting [29]. Various reviews have questioned their ability to generalize and function viably in practice [23, 29]. Moreover, the fact that histological and gene expression based approaches require invasive biopsies and only capture data from a small portion of a potentially heterogeneous tumor has raised significant concerns [30, 31].

4. Image Processing and Deep Learning:

Recently, radiomics has attempted to address some of these concerns. There is significant evidence demonstrating that radiomic methods can noninvasively find prognostic signatures that also capture intratumor heterogeneity [30–33]. Furthermore, radiographic data has been shown to capture tumor characteristics that are both consistent with immunohistochemical analysis [32, 33] and have also been able to define signatures that are associated with underlying gene-expression patterns [30, 33]. These findings demonstrates that image processing and deep learning have the potential to synthesize patterns found in
other techniques and improve prognostication. While few studies that have attempted to do so, the ones that have used these methods have had promising results [15, 34].

However, a significant barrier to the continued development of recurrence prediction using image processing and deep learning has been the lack of available, annotated datasets. As is evident from the LIDC-IDRI dataset and nodule segmentation, the availability of a well documented and extensive dataset can rapidly accelerate the development of technology within the image processing and deep learning domain. However, the task of recurrence prediction requires a dataset with both images and a set of many specific clinical features. At a minimum, these features include treatment data to determine the procedure used (surgical vs. radiotherapy, vs. chemotherapy, etc.), progression data to determine the incidence of recurrence (which also requires a sufficient follow up period), and usually, if no radiologist is available, data to localize the tumor within the image. The NLST dataset provides all of these features except for localization data. However, our fully automated pipeline enables us to circumvent this requirement.

4.3 Proposed Method

The recurrence prediction framework we present consists of several intermediate steps (see Figure 4.1). First, the predicted segmentation from the previous stage is used to generate a 3D ROI around the identified tumor. Next, this ROI is passed as input to a recurrence risk prediction CNN. Finally, the CNN’s output, clinical data, and geometric features of the tumor, which are extracted from the previous stage’s predicted segmentation, are used as features in a random forest classifier. Each of these steps is discussed in further detail below.
4.3.1 ROI Generation

![ROI Generation Diagram]

**Figure 4.3: ROI generation.** Before training our recurrence prediction network, we first generate ROIs around the tumors using the predicted segmentations from the previous stage in our system. This is done by resampling everything to $1 \times 1 \times 1$ mm world coordinates and extracting a $50 \times 50 \times 50$ ROI centered around the predicted segmentation.

The first step in ROI generation is converting from voxels to world coordinates (mm). That is, the raw 3D volume is resampled so that each data point represents a $1 \times 1 \times 1$ mm cube of real world data. Although it ranges, a typical raw CT series will have an axial slice thickness ($z$-direction) of about 2.5 mm and a pixel spacing ($x$-$y$ direction) of about 0.75 mm. (The exact values are retrieved from the metadata provided in the images’ DICOM headers). Converting this data to world coordinates is necessary for two key reasons. The first is to have a consistent amount of data represented in each of the $x$, $y$, and $z$ directions respectively. The second is to have a consistent scale between each of the volumes, which is crucial now that the task at hand is classification (in which the relative size of the tumor plays role).

After resampling the volume, a ROI can be extracted. First, the predicted segmentation is also resampled to meet the dimensions of the new image in world coordinates. Then, the $x$-$y$ coordinates of the ROI are set as a $50 \times 50$ mm box centered around the predicted tumor segmentation. This same 2D ROI is then extracted from (max) 25
slices above and (max) 25 slices below the main slice to generate the 3D 50 x 50 x 50
mm ROI.\textsuperscript{2} For nodules that are at the extremity of the lungs (one of the first or last
slices in the series), 25 slices can not always be taken from both above and below the
image. Therefore, if the main slice is at least 20 slices from the end of the image,
20 slices are taken from the end closer to the extremity and an excess of 5 is taken
in the other direction. If the main slice is closer than 20 slices from the end, the
volume is not used. After this process, all extracted volumes are passed as input to
the recurrence prediction network.

4.3.2 Proposed Network Architecture

The CNN architecture we use for recurrence prediction is illustrated in Figure 4.4.
It follows a typical CNN structure similar to the one used by Honsy et al. [15]. It
contains four 3D (3 x 3 x 3) convolutional layers with 32, 64, 128, and 256 filters,
respectively. Each convolutional layer is followed by a LeakyReLU activation layer

\textsuperscript{2}The “main” slice refers the extracted axial slice upon which the segmentation was originally generated.
that has an alpha value of 0.1. Additionally, the 2nd and 4th convolutional units are followed by \((3 \times 3 \times 3)\) Max Pooling layers. Following the 4 convolutional units, the images are flattened and fed into 4 fully connected layers with 1,024, 512, 256, and 2 units, respectively. Each fully connected layer has a dropout of 25%. Finally, a sigmoid activation function is applied to get the final prediction.

The remaining hyperparameters of our network are similar to those of our segmentation stage. For the optimizer, we use the Adam optimizer with a global learning rate of 0.0001. We also use weighted binary cross-entropy (see Equation 3.3) for our network’s loss function with a weight of 3.0 for the positive class. Our choice is motivated by the disproportional number of patients who do not experience recurrence to those who do experience recurrence. There are about 3 times as many non-recurrent patients as recurrent patients. Once again, the model is implemented using the Keras 2.2.4 and tensorflow 1.14.0 as the backend.

4.3.3 Training

Prior to training the network, the images are partitioned into training, testing, and validation sets. Since many patients have multiple nodules, we first create a random split based off patient ids. To ensure an even class distribution, we stratify this split on recurrence status. We then partition the corresponding images for each patient accordingly. The final breakdown is 620 images in the training set, 72 in the validation set, and 177 in the test set.

Due to this limited number of images, data augmentation once again plays a critical role. Similar to the segmentation network, all augmentation is done in real time during training instead of as part of preprocessing. Augmentation operations include random brightness shifts, vertical shifts, horizontal shifts, vertical flips, and horizontal flips. Figure 4.5 illustrates an example of the resulting image after each
Figure 4.5: Recurrence prediction network data augmentation. An example of the data augmentation process for a $50 \times 50 \times 50$ input cube to the recurrence prediction network. Note, each image shown above illustrates only one of the possible results from the augmentation process. The magnitude of each transformation is not fixed, but rather chosen randomly during training. Likewise, the choice of whether or not to apply the horizontal and vertical flips is also chosen randomly. After all augmentation transformations are applied, the image is normalized to the 0-1 range and input to the network.

4.3.4 Integrating Clinical and Geometric Data With a Random Forest Classifier

The final part of our system uses a random forest classifier to combine our neural network’s predictions with clinical data and geometric tumor features that are extracted from the segmentation predicted in the previous stage. The model’s parameters (number of estimators and maximum tree depth) are selected using cross validation. An in depth discussion of the features selected is provided below, and an illustration of the random forest pipeline is shown in Figure 4.6.

Unlike previously, the incorporation of clinical and geometric data means that we are now dealing with features that are both on a per patient basis (e.g. cancer stage, age, etc.) and a per nodule basis (network prediction, geometry, etc.). As a result, we are left with the choice to either (a) replicate a patient’s clinical features for each of their nodules and classify on a per nodule basis, or (b) create a scheme to condense...
Figure 4.6: Integrating clinical and geometric data with a random forest classifier. The final part of our system uses a random forest classifier to combine our neural network’s predictions with clinical data and geometric tumor features, which are extracted from the predicted segmentation. The output of this model is the final prediction of our system.

the nodule information, get a single value for each nodule feature per patient, and classify on a per patient basis. With option (a) there is a significant risk of creating false correlations between variables, which is a particular issue given the limited size of our dataset. For example, if there is a patient with two nodules and images taken from all three years of the study, that patient’s exact combination of clinical features will exist in six different samples. On the other hand, the combination of clinical features from a patient with one nodule and images from one year will only exist in one sample. Therefore, we choose to go with option (b). A list of the features used is provided below.
1. **Clinical Features:**

The clinical features used are either extracted directly from the NLST dataset or inferred from the other attributes. (See the section on NLST feature engineering in Chapter 2.2 for more details on inferred features). The features included are:

(a) *Cancer stage*: cancer stage on the I-IV stage range

(b) *Gender*

(c) *Age*

(d) *Adjuvant therapy*: whether adjuvant therapy was administered after surgery

2. **Geometric features:**

Geometric features are extracted from the 2D segmentation predicted in the previous stage. If a patient has more than one tumor, we extract data from their largest tumor. The features extracted are:

(a) *Diameter*: largest cross-sectional diameter of the tumor (mm)

(b) *Average attenuation*: average intensity of the tumor’s pixel values in the original image

(c) *Perimeter*: perimeter of the tumor (mm)

3. **Neural Network Predictions:**

Given our aforementioned decision to classify based off patients rather than nodules, when a patient has more than one nodule, we have to transform the set of the network’s predictions on each nodule, $\hat{Y}_{nodule}$, into a single probability, $\hat{y}_{patient}$. We experiment with two different methods for this transformation:

(a) *Mean Prediction*: the average predicted probability of all of the patient’s
nODULES is used as the predicted probability for the patient:

\[
\hat{y}_{\text{patient}} = \frac{1}{|\hat{Y}_{\text{nodules}}|} \sum_{y \in \hat{Y}_{\text{nodules}}} y
\]  

(4.1)

(b) **Most Confident Prediction**: the “most confident” predicted probability of any of the patient’s nodules is used as the predicted probability for the patient:

\[
\hat{y}_{\text{patient}} = \hat{y}_i \in \hat{Y}_{\text{nodules}}
\]

s.t.

\[
i = \arg \max_j (|0.5 - \hat{Y}_{\text{nodules},j}|)
\]  

(4.2)

Only one of these methods is used to determine the value of the final neural network prediction (rather than using both as features in the random forest model). Therefore, we evaluate the standalone performance of each. Ultimately, as is discussed further below, the most confident prediction is chosen over mean prediction. The results are provided in the following section.

### 4.4 Results and Experiments

#### 4.4.1 Evaluation Metrics

We use several different metrics to evaluate the performance of our model:

1. **Precision**: (see Equation 3.8)

2. **Recall**: (see Equation 3.9)

3. **ROC-AUC**: the ROC curve (receiver operating characteristic curve) plots classification performance at all thresholds. Specifically, at each threshold, it plots...
the true positive rate (i.e. recall) against the false positive rate (i.e. the ratio of true positives to false positives and true negatives). AUC (area under the curve) measures the 2D area underneath the ROC curve. It provides an aggregate metric for performance at different thresholds and characterizes a model’s ability to discriminate between classes.

In addition to measuring performance on the overall test set, we also evaluate our model on subgroups of patients based on cancer stage. The metrics used in each subgroup are chosen to reflect our model’s capability to identify (a) high risk stage I patients and (b) low risk stage II patients. As is discussed in our review of the TNM staging system in Section 4.1.2, these are particular prognostic needs that have been expressed in previous medical literature [29]. For each group, we measure the following:

1. All patients: accuracy, ROC-AUC, precision, recall
2. Stage I patients: accuracy, recall
3. Stage II patients: accuracy, precision

It is worth mentioning that the limited size of our dataset poses certain limitations for our evaluation. Specifically, evaluating staging subgroups individually means that some of our testing subsets are relatively small. This makes it harder to guarantee that the same results generalize to larger populations. Nonetheless, we still choose to include the results for these subgroups because of the aforementioned medical significance.

4.4.2 Results

We evaluate our model at two separate stages – once after the neural network predictions and a second time after the random forest predictions. At each stage, we run
a number of experiments to compare different methods and benchmark performance. The results are provided in the following section along with a further discussion.

**Neural Network Performance**

Before evaluating the overall performance of our system with the random forest model, we evaluate the performance of just the neural network. We first test the network on all the nodules in the test set. We then test it on a per patient basis using the mean nodule prediction and most confident nodule prediction methods mentioned above. The results are shown in Tables 4.1 and 4.2 and Figures 4.7 and 4.8. Based on these findings, we choose to use most confident prediction in the random forest over mean prediction. Our choice is motivated by its superior AUC and precision.

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
<th>AUC</th>
<th>Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>All nodules</td>
<td>73%</td>
<td>60%</td>
<td>18%</td>
<td>30%</td>
</tr>
<tr>
<td>Patients (w/ mean nodule prediction)</td>
<td>71%</td>
<td>62%</td>
<td>17%</td>
<td>38%</td>
</tr>
<tr>
<td>Patients (w/ most confident nodule prediction)</td>
<td>69%</td>
<td>65%</td>
<td>17%</td>
<td>43%</td>
</tr>
</tbody>
</table>

**Table 4.1: Neural network overall recurrence prediction performance.** Network performance on all nodules, on patients using mean nodule prediction, and on patients using most confident nodule prediction.

<table>
<thead>
<tr>
<th>Method</th>
<th>Stage I Accuracy</th>
<th>Stage I Recall</th>
<th>Stage II Accuracy</th>
<th>Stage II Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>All nodules</td>
<td>75%</td>
<td>12%</td>
<td>60%</td>
<td>50%</td>
</tr>
<tr>
<td>Patients (w/ mean nodule prediction)</td>
<td>75%</td>
<td>17%</td>
<td>75%</td>
<td>undefined</td>
</tr>
<tr>
<td>Patients (w/ most confident nodule prediction)</td>
<td>73%</td>
<td>17%</td>
<td>75%</td>
<td>undefined</td>
</tr>
</tbody>
</table>

**Table 4.2: Neural network recurrence prediction performance by cancer stage.** Network performance by stage on all nodules, on patients using mean nodule prediction, and on patients using most confident nodule prediction.
Overall System Performance

After evaluating the network, we evaluate the performance of the overall system with the random forest classifier. To demonstrate the significance of our results, we also
compare the performance of our final model to a random forest that uses only staging and other relevant clinical data (i.e. no geometric or network prediction features). The results from our model and this comparison are shown in Tables 4.3 and 4.4 and Figure 4.9.

![Figure 4.9: Random forest ROC-AUC for recurrence prediction. (a) ROC-AUC of the random forest model using only staging and clinical data. (b) ROC-AUC of the proposed random forest model.](image)

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
<th>AUC</th>
<th>Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Staging and Clinical Random Forest Model</td>
<td>71%</td>
<td>56%</td>
<td>16%</td>
<td>42%</td>
</tr>
<tr>
<td>Proposed Random Forest Model</td>
<td>78%</td>
<td>73%</td>
<td>44%</td>
<td>66%</td>
</tr>
</tbody>
</table>

Table 4.3: Random forest overall recurrence prediction performance

<table>
<thead>
<tr>
<th>Method</th>
<th>Stage I Accuracy</th>
<th>Stage I Recall</th>
<th>Stage II Accuracy</th>
<th>Stage II Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Staging and Clinical Random Forest Model</td>
<td>75%</td>
<td>0%</td>
<td>25%</td>
<td>0%</td>
</tr>
<tr>
<td>Proposed Random Forest Model</td>
<td>75%</td>
<td>17%</td>
<td>100%</td>
<td>100%</td>
</tr>
</tbody>
</table>

Table 4.4: Random forest recurrence prediction performance by cancer stage
For each of the models tested, we also include the feature importances. The importance \( f_i \) of the \( i \)'th feature in a decision tree is defined by the following:

\[
\begin{align*}
    n_j &= |X_j| \cdot C_j - |X_{\text{left}(j)}| \cdot C_{\text{left}(j)} - |X_{\text{right}(j)}| \cdot C_{\text{right}(j)} \\
    f_i &= \frac{\sum n_j \text{ split on } i}{\sum n_k}
\end{align*}
\] (4.3)

where \( n_j \) is the importance of the \( j \)'th node in the tree, \( X_j \) is the set of samples reaching the \( j \)'th node, and \( C_j \) is the impurity value of \( j \)'th node.\(^3\) The importance of a feature in a random forest is the average of its importance in each tree.

![Feature importances for Staging and Clinical Data Model](image1)

**Figure 4.10:** Feature importances of staging and clinical random forest model

![Feature importances for Proposed Model](image2)

**Figure 4.11:** Feature importances of proposed random forest model

\(^3\)The impurity of a node is determined by the homogeneity of the samples reaching the node.
4.5 Discussion

The TNM staging system demonstrates the ability to capture a certain, but limited amount of information about a patient’s risk for recurrence. It is in fact statistically proven that patients in later stages are more likely to experience recurrent cancer (see Figure 4.2). It is also shown from the staging and clinical data model in our experiments that these factors can be utilized to make non-trivial predictions. However, prognostic tools that depend on this data exclusively are also inevitably insufficient. Putting excessive weight on whether or not the patient has progressed to a later stage (typically considered progression to stage II or further), has unavoidable limitations. Specifically, it often ignores high risk early stage patients and inflates the risk of later stage patients. These issues are evident from clinical studies [29] as well as our own experiments. Our model that uses staging and clinical data exclusively is completely unable to classify stage I patients that experienced recurrence or stage II patients that did not. Interestingly, even when other clinical features like age are considered a primary factor (see Figure 4.11), these patterns persist.\textsuperscript{4}

Nonetheless, our neural network also proves to have its own drawbacks when evaluated on its own. Most prominently, when the predictions are a combined on a per patient basis, it has the opposite issue as the staging and clinical data model - all stage II patients are classified as non-recurrent (0% recall and undefined precision). While this issue may not be directly related to the fact that the patients are in stage II, but rather an unfortunate consequence of the model’s low overall recall and the limited size of our test set, it can still be argued that the model should be more sensitive to these patients. Moreover, the network’s overall performance is not significantly better than that of the staging and clinical model. Nonetheless, these

\textsuperscript{4}The staging and clinical data model used in our experiments should only be interpreted as (a) a single example of how this data can be insufficient and (b) a benchmark for our results. It is not what is used in clinical practice nor is it meant to be representative of the clinical prognostic process.
results are still significant as they reinforce the fact that deep learning can produce non-trivial predictions for risk of recurrence and pick up on critical radiographic patterns. Consequently, as our results demonstrate, synthesizing these findings with other forms of data can create a combined model that mitigates the issues of each of the individual models.

The results of our final model demonstrate the efficacy of our approach and more generally, the potential of deep learning to improve recurrence risk prognostication. By combining neural network predictions with extracted geometric features and clinical and staging data, the model outperforms all other approaches tested. Both overall performance, and performance within staging subgroups are significantly increased. What’s more, as is evident from the feature importances in Figure 4.11, our neural network, the extracted tumor geometry features, and clinical data all play a crucial role in the model’s final decision. These findings are significant for our study, as the integration of the three different types of data is unique to our approach. The extraction of geometric features from the tumors and use of the extensive clinical data from the NLST dataset is only possible with our fully automated segmentation pipeline. Given the novelty of not only our system, but also deep learning for recurrence risk prognostication in general, these results are promising.
Chapter 5

Conclusion

5.1 Overview

In this paper we present a fully automated system for the segmentation and recurrence risk prediction of surgically resected lung tumors. In recent years, deep learning models, and CNNs in particular, have become the state of the art in lung nodule segmentation. Within the realm of recurrence prediction however, these types of approaches have just recently arisen. A significant barrier to the continued development of such models is the lack of viable datasets. Using any sort of image processing method to address recurrence prediction requires a significant amount of very specific features, many of which are not available in standard lung cancer imaging collections. These features include treatment data to determine the procedure used (surgical resection vs. radiotherapy, vs. chemotherapy, etc.), progression data to determine recurrence status (which also requires a sufficient follow up period), and usually, data to localize the tumor within the image when no radiologist is available. However, by taking advantage of the progress in nodule segmentation, our proposed system circumvents this need for localization data. While, certain CAD methods for malignancy classification have also employed fully automated segmentation to classification pipelines, to the best of our knowledge, our proposed system is
the first to do this for recurrence prediction.

## 5.2 Contributions

Our work makes a significant contribution to lung cancer radiomics by demonstrating that segmentation algorithms trained on LIDC-IDRI can generalize to other datasets and function in multitask pipelines. While previous multitask models (primarily for malignancy classification) have been successfully implemented [9, 13], many have used LIDC-IDRI exclusively. As a result, their ability to generalize remains largely untested and their set of possible classification tasks is restricted by the data available in LIDC-IDRI. Our study addresses both of these issues. First, although we do not exhaustively test our segmentation model on an independent dataset, we do validate its results by testing its relative accuracy on SPIE-AAPM and also by demonstrating its functionality in our full pipeline. Second, by using the NLST dataset and tackling recurrence prediction, our work significantly expands the tasks possible with automated segmentation.

Furthermore, our work also makes a significant contribution to deep learning recurrence prediction research, as well as the general search for a better recurrence prognostication tool. Our fully automated pipeline enables the use of the NLST dataset and its expansive set of both imaging and non-imaging data. This has several implications. First off, it introduces a dataset to deep learning methods for recurrence prediction that has never been used. Secondly, it enables us to have better selection criteria. Specifically, access to post-surgical results allows us to make sure we exclusively use patients with completely resected tumors and do not consider those who have with residual disease after surgery. As we discuss previously, residual disease makes it unclear if a patient can ever be considered “cancer free” and is also highly correlated with recurrence [29]. Therefore, since the images used are taken before
surgery, including patients with residual disease can skew predictions. Additionally, access to disease progression data, as opposed to simply survival data, means that we can predict true disease free survival. Often, the outcome variable in recurrence prediction models is simply based off of survival time where the event of death is often, but not always, due to recurrent cancer. Lastly, using the NLST dataset allows us to integrate crucial clinical features into our final model. Coupled with the geometric features extracted after segmentation, this enables the proposed model to take advantage of multiple sources of information. Being able to use factors like cancer stage, age, gender, and whether or not post-surgical adjuvant therapy was administered improves performance and also increases the interpretability of our model. While beyond the scope of our study, it also introduces the possibility to explore how these factors influence predictions and the subsequent implications for clinical practice. Evidently, the novelty of our proposed system leaves substantial room for future work.

5.3 Limitations and Future Work

While the findings of our work are significant, additional studies that explore ways to refine our approach further could make major contributions. Deep learning, radiomics, and lung cancer research are all constantly evolving fields. Each stage in our pipeline will have the continual opportunity to be improved by novel approaches in the future.

For our segmentation stage, there are multiple directions that can be pursued to improve our pipeline. One logical next step would be to explore 3D segmentation. While 2D segmentation was sufficient for the purposes of our study, surveys have found that 3D segmentation is often more effective than 2D [67].\(^1\) Given the expansion to

\(^1\)Many of the works demonstrating the efficacy of 3D segmentation use a cropped ROI around
3D in the recurrence stage of our system, 3D segmentation would integrate into our pipeline particularly well. However, it is worth noting that extending our proposed architecture to 3D would require a significant amount of GPU memory. Unlike the recurrence prediction stage, which takes an ROI as input, the input to segmentation is the entire image. In 3D, this will be enormous. Nonetheless, if these resources are available, 3D segmentation would be a worthwhile pursuit. Otherwise, another potential direction for improvement is expanding segmentation to include image wise classification. Currently, our system takes advantage of the fact that many lung cancer CT image collections, like SPIE-AAPM and NLST for example, designate which cross sectional slices contain tumors. A model that could autonomously determine whether or not there is a tumor present and then segment the image accordingly would take our segmentation stage a step further. One option would be to create a separate detection network and then segment the positive images. However, in the interest of avoiding an excessively complex system, a joint learning method would be more desirable. One potential approach is to use information density. In [68], Hsu et al. presents a framework that uses a trimmed information density estimator to quantify information leaking features. While originally developed for the identification and obfuscation of information leaking features (i.e. pixels or groups of pixels that reveal the label of an image), the same framework can be used for the opposite purpose. That is, in our case, determining which areas of an image reveal the presence of a tumor can be used as a nodule segmentation mechanism. Implementing this framework into our pipeline successfully would not only improve our system, but could also have significant implications for radiomics and CAD research in general.

There is also significant room for future work within our recurrence prediction stage. For example, creating a model that predicts not just risk of recurrence, but also the potential efficacy of adjuvant therapy, could contribute significantly to the nodule to reduce the problem space. Therefore, issues surrounding GPU memory are not as pertinent.
medical field. Additionally, integrating other forms of data, like histological or gene expression features could yield even more promising results. Unfortunately, a good amount of this work is also beyond our domain. Synthesizing the findings of our study with clinical research or research in medical treatments requires a significant amount of medical experience. Our position outside of the medical field is an acknowledged limitation of this thesis.

Lastly, an additional limitation of our study is the small sample size used. While working with limited datasets is common in medical imaging and can be partially handled with data augmentation, acquiring a larger or additional dataset, particularly for training and testing our recurrence prediction model, would be extremely useful. It would not only likely improve performance, but also significantly contribute to demonstrating our system’s ability to generalize.

5.4 Concluding Remarks

Lung cancer remains the leading cause of cancer related mortality across the globe. While advances in screening and diagnostic technology have enabled significant improvements in detection and diagnosis, the disease’s five year survival rate is still startlingly low. High rates of tumor recurrence is one of the main factors contributing to this statistic. Thus, there is a strong need to improve the methods that are used to predict a patient’s risk for recurrence. Our work demonstrates that first, state of the art nodule segmentation technology can generalize and enable automated pipelines for a wide range of multitask systems like recurrence prediction, and second, that integrating deep learning and image processing methods into recurrence prediction systems has the potential to improve current prognostic tools. To the best of our knowledge, our proposed framework is the first fully automated segmentation and recurrence risk prediction system to be implemented.
References


APPENDICES

A  Abbreviations

CAD    Computer Aided Detection
CADx   Computer Aided Diagnosis
CNN    Convolutional Neural Network
CT     Computed Tomography
EM     Expectation Maximization Algorithm
FCM    Fuzzy C-Means Algorithm
FDA    Food and Drug Administration
FNIH   Foundation for the National Institutes of Health
GBD    Global Burden of Disease
IHC    Immunohistochemistry
IHME   Institute of Health Metrics and Evaluation
LUNA16 2016 Lung Nodule Analysis Challenge
LIDC-IDRI Lung Image Database Consortium and Image Database Resource Initiative (a dataset)
MRI    Magnetic Resonance Imaging
NCI    National Cancer Institute
NLST   National Lung Screening Trial (a dataset)
ROI    Region of Interest
SPIE-AAPM Society of Photo-Optical Instrumentation Engineers and American Association of Physicists in Medicine (a dataset)
TNM    Tumor, Node, Metastasis – the convention used for cancer staging
B Code Availability

The code used to implement all models and experiments in this thesis is included in the following repository: https://github.com/maggiebasta/lung-cancer-thesis.