THE UNIVERSITY OF CHICAGO

IMPLEMENTATION AND ANALYSIS OF ARTIFICIAL INTELLIGENCE FOR PLEURAL MESOTHELIOMA ON COMPUTED TOMOGRAPHY SCANS AND COVID-19 ON CHEST RADIOGRAPHS

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES
AND THE PRITZKER SCHOOL OF MEDICINE
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

COMMITTEE ON MEDICAL PHYSICS

BY
MENA SHENOUDA

CHICAGO, ILLINOIS
JUNE 2024

"*Deep in the human unconscious is a pervasive need for a logical universe that makes sense. But the real universe is always one step beyond logic.*"

– Frank Herbert

TABLE OF CONTENTS

LIST OF FIGURES vii

LIST OF TABLES xii

ACKNOWLEDGMENTS xiv

ABSTRACT xvii

1 INTRODUCTION 1
1.1 Artificial Intelligence 1
1.2 Machine Learning 2
1.2.1 Cross-validation 6
1.3 Deep Learning 7
1.3.1 Convolutional neural networks 10
1.3.2 Model calibration 13
1.4 Texture Feature Analysis 16
1.5 The Role of AI in Medicine 17
1.6 Clinical Pipeline, Characterization, and Potential for AI in the Assessment of Mesothelioma 18
1.7 Clinical Pipeline, Characterization, and Potential for AI in the Assessment of COVID-19 21
1.8 Outline of Dissertation 22

2 CONVOLUTIONAL NEURAL NETWORKS FOR SEGMENTATION OF PLEURAL MESOTHELIOMA: ANALYSIS OF PROBABILITY MAP THRESHOLDS 24
2.1 Introduction 24
2.2 Methods 26
2.2.1 Patient population 26
2.2.2 Model calibration 28
2.2.3 Tumor volume and Dice similarity coefficient 29
2.2.4 Statistical methods 30
2.3 Results 31
2.3.1 Tumor volume and DSC 31
2.3.2 Model calibration using TS 42
2.4 Discussion 43
2.5 Conclusion 47

3 RADIOMICS FOR DIFFERENTIATION OF SOMATIC BAP1 MUTATION ON CT SCANS OF PATIENTS WITH PLEURAL MESOTHELIOMA 49
3.1 Introduction 49
3.2 Methods 50
3.2.1 Patient selection and sample collection 51

3.2.2 Image data curation and segmentation . . . 51
3.2.3 Image resampling and gray-level discretization . . . 52
3.2.4 Feature extraction . . . 54
3.2.5 Data imbalance . . . 55
3.2.6 Machine learning model and feature selection . . . 55
3.2.7 Evaluation metric and statistical analysis . . . 57
3.3 Results . . . 57
3.3.1 Tumor volume . . . 57
3.3.2 Classification performance . . . 58
3.3.3 Change of k-fold and number of features . . . 60
3.3.4 DSC and classification performance of unmodified segmentations . . . 61
3.4 Discussion . . . 62
3.5 Conclusion . . . 66

4 ASSESSMENT OF A PRE-TRAINED DEEP LEARNING MODEL FOR COVID-19 CLASSIFICATION ON CXRS . . . 67
4.1 Introduction . . . 67
4.2 Methods . . . 68
4.2.1 Datasets . . . 68
4.2.2 Image preprocessing . . . 70
4.2.3 Model implemented . . . 72
4.2.4 Statistical analysis . . . 73
4.3 Results . . . 76
4.3.1 Current test set only . . . 76
4.3.2 Original test set versus current test set . . . 78
4.4 Discussion . . . 88
4.5 Conclusion . . . 92

5 CONCLUSIONS AND FUTURE DIRECTIONS . . . 94

APPENDIX A IMPACT OF MODEL RETRAINING ON A DEEP LEARNING MODEL IN THE TASK OF COVID-19 CLASSIFICATION ON CXRS: A PILOT STUDY 99
A.1 Introduction . . . 99
A.2 Methods . . . 101
A.2.1 Datasets . . . 101
A.2.2 Image Preprocessing . . . 102
A.2.3 Model Training Scheme . . . 102
A.2.4 Analyses and Comparisons . . . 103
A.3 Results . . . 107
A.3.1 Experiment I: Recalculating phase 3 weights . . . 107
A.3.2 Experiment II: Fine-tuning phase 3 weights . . . 107
A.3.3 Experiment III: *L2* regularization . . . 108
A.3.4 Experiment IV: Recalculating phase 3 weights after repartitioning . . . 108
A.3.5 MIDRC Grand Challenge . . . 111

A.4 Discussion 111
A.5 Conclusion 114

REFERENCES 115

LIST OF FIGURES

1.1 Visual representation of AI, ML, and DL, where each field is a subset of the prior. 2

1.2 Geometry of least squares. The schema displays the columns of dataset X, which span the blue hyperplane. The truth label vector is projected on this hyperplane. The distance between the weights resulting from the projection ($\vec{w}$) and nonoptimal weights ($\tilde{\vec{w}}$) due to a residual is also shown. 3

1.3 Visualization of gradient descent with (a) displaying a 3D convex function that intersects a hyperplane, where the intersection shows the lowest value the loss function can take. The gradient descent process can be visualized in 2D in (b), with w^* showing the lowest value of the convex loss function $f(\vec{w})$. 5

1.4 Schema depicting k-fold CV. For leave-one-out CV, k would equal the number of samples n, and $n-1$ samples would be used for training, with the n-th sample used for testing. 7

1.5 An example of a neural network with one hidden layer in blue. The values $x_0 = 1$ and $h_0 = 1$ are the bias terms used in the model. Of note, this figure depicts "fully connected layers" as every input neuron is connected to every output neuron. 8

1.6 Common nonlinear activation functions used in DL algorithms. 9

1.7 Demonstration of a 2D vertical edge detection convolution filter operating on an input feature map. The filter was applied with stride 1, as the second row in the figure displays the movement across one pixel when compared to the top row. Further, no zero-padding was applied to the input feature map, which resulted in a "valid" convolution. A valid convolution is defined as a convolution only performed over pixels where the convolution filter overlaps completely with the input feature map—values outside the filter have no effect on the output feature map. The bottom row of the figure displays the output feature map after the convolution filter is applied to the entirety of the input feature map. 12

1.8 Probability maps for two potential classes ("tumor" and "no tumor") in medical image semantic segmentation tasks. These values would be the result of the logits input to the final layer, i.e., activation function, of the model. Because this is a binary task, the pixel-wise values across the two class channels will add to 1. The argmax function could be used to return the final binary segmentation. 13

1.9 Linear measurements made by a radiologist to quantify tumor burden. Using longitudinal summations of these measurements, patient response is evaluated and the efficacy of treatment is assessed. 19

2.1 Schema demonstrating the methodology employed. Beginning from left to right: (a) 2D CT sections of a patient were input to the (b) VGG16/U-Net, and (c) the probability maps were generated. The probability maps were binarized using a range of thresholds, where (d) the reference standard was provided by a radiologist by modifying the generated segmentations at the 0.5 threshold. Lastly, the reference standard was compared to (e) the probability maps binarized at the various thresholds, using the percent difference of volume and DSC as the two figures of merit. U-Net figure reprinted, with permission, from [74]. 30

2.2 Differing contours on the same section of the same patient created with an adjustment in the CNN probability threshold. Purple represents the radiologist reference outline, and green represents the CNN pixel-wise segmentation prediction of tumor with (a) a probability threshold of 0.5 (average DSC over all sections: 0.357) and (b) a probability threshold of 0.001 (average DSC over all sections: 0.476). 32
2.3 Boxplots showing the DSC values obtained for tumor comparisons acquired between the radiologist and the deep CNN at six different thresholds. The solid red lines display the median DSC value at each probability threshold. The dashed red line displays an average human interobserver DSC of 0.74 achieved between radiologists in the task of segmenting mesothelioma on CT scans from a separate dataset [42]. 33
2.5 Scatter plot displaying the correlation between the tumor volumes calculated from the CNN contours obtained at the 0.5 threshold and the radiologist reference contours. One outlier at (14.2×10^5 mm^3, 33.1 $\times10^5$ mm^3) is not shown. The dashed red line represents the identity line. 34
2.4 Matrix of p-values when comparing the absolute percent difference of volume (a) and DSC (b) across thresholds. Red indicates a significant difference ($p < 0.0033$ after Bonferroni correction), and green indicates a failure to achieve significance, as determined by the Wilcoxon signed-rank test. 35
2.6 Bland-Altman plot of the relative differences between reference and CNN-based tumor volumes at the 0.5 threshold. The red band highlights differences within 5% of 0. 36
2.7 Histogram of the CNN output thresholds that maximize DSC and minimize percent difference of volume. 37
2.8 The average absolute percent difference of volume (and its minimum) and the average DSC (and its maximum) across all cases for the entire threshold range. . 37
2.9 Example images from four of the six scans that exceeded the 95% agreement limits as shown in the Bland-Altman plot (Figure 2.6). Yellow arrows point to regions where the CNN predicted tumor at the 0.5 threshold. Purple outlines are the radiologist reference contours. 38
2.10 Boxplots showing the DSC values obtained for the subset tumor comparisons acquired between the radiologist and the deep CNN at six different thresholds. The solid red lines display the median DSC value at each probability threshold. The dashed red line displays an average human interobserver DSC of 0.74 achieved between radiologists in the task of segmenting mesothelioma on CT scans from a separate dataset [42]. 39
2.11 Matrix of p-values when comparing the absolute percent difference of volume (a) and DSC (b) across thresholds for the subset analysis. Red indicates a significant difference ($p < 0.0033$ after Bonferroni correction), and green indicates a failure to achieve significance, as determined by the Wilcoxon signed-rank test. 41

2.12 P-values comparing the absolute percent difference of volume and DSC volumes between the entire scan and the subset sections selected. Red indicates a significant difference, and green indicates a failure to achieve significance, as determined by the Wilcoxon signed-rank test. Significance was achieved at $p = 0.0083$, after correcting for six comparisons. 42
2.13 Example of a tumor plus effusion case in the left hemithorax at the various stages of post-processing to calculate the temperature. The bottom right image is the calibrated probability vector for the "disease" class, which is the output of the sigmoid activation function of the logit vector $\boldsymbol{z}_i$ scaled to temperature $T = 3.4$. 43

3.1 Pipeline incorporated in this study, beginning with the patient cohort curated and ending with the machine learning models used for the *BAP1* classification task. 50
3.2 Histogram of the (a) pixel spacing and (b) slice thickness of CT sections of the original 149 scans. The red vertical line depicts the mean value in each of the distributions to which resampling was performed. 54
3.3 Histogram of the tumor volume categorized by *BAP1* mutation status. The difference in tumor volume between wild-type and mutated tumors failed to achieve statistical significance. 58
3.4 (a) ROC curves depicting the true-positive and false-positive fractions of the top-three performing classifiers in the task of differentiating somatic *BAP1* mutation status using feature values extracted from segmented regions. ROC curves were fitted using software created by Metz et al. [109]. (b) Distributions of the decision tree classifier prediction scores across all cases. The histograms were normalized to have equal area of one. 59

4.1 The image preprocessing pipeline used throughout this work. Top panel: a patient's standard CXR was resized to 256 × 256 pixels and the lung region was subsequently segmented and cropped, which generated a rectangular region containing only the lung (the small lung region). Bottom panel: after the segmentation task, the cropping was performed in the same dimensional space as the original image. U-Net figure reprinted, with permission, from [74]. DenseNet-121 figure copyright 2017 IEEE [131]. 72
4.2 Box and whisker plot of the severity score distribution of all 1,972 cases of the original test set. Ten cases were randomly chosen from each part of the box and whisker plot to ensure an equal representation of cases from all possible PXS scores assigned, totaling 50 random cases selected from the original test set to determine the robustness of the PXS scores. 75
4.3 ROC curves for the classification of COVID-positive and COVID-negative patients based on CXRs using the current test set. No additional training or validation was performed. ROC curves were fitted using software created by Metz et al. [109]. 76

4.4 Comparison of ROC curves and AUC values between (a) original and (b) current test sets. AUC values for the three classification algorithms were consistently lower for the current test set compared with the original. Figure 4.4a is the same figure as Figure 6 in ref. [44] (reprinted with permission) and Figure 4.4b is the same as Figure 4.3. 79
4.5 Histogram of the imaging exam dates, categorized by the current test set, the original training set, and the original test set. The current test set had a much larger date range, spanning March 15, 2020 to January 1, 2022. 80
4.6 (a) Scatter plot of a subset of images from the original test set that displays the mRALE score determined by the radiologist and the COVID severity as determined by the DL model described by Li et al. [138]. The regression line is shown in blue where the shaded blue region is the 95% CI. (b) Bland-Altman plot displaying the agreement between the methods of assessing COVID severity. The outlier outside the 95% limits of agreement demonstrated a collapsed left lung with possible effusion [127]. 82
4.7 Bar plot depicting the resulting AUC values when controlling for PXS scores using the PXS score bin edges defined in Table 4.7. The 95% CIs were calculated by bootstrapping the AUC values 2000 times. 83
4.8 Histograms of the severity scores calculated from the original (a) and current (b) test sets. Both distributions demonstrate a strong right skew with higher frequency of positive cases having larger PXS scores. 83
4.9 Histogram of the prediction scores of the DL model for both test sets. The distribution (a) before February 3, 2021, and (b) the entire data range. February 3, 2021 was chosen as the cutoff date as that is the last date which had an overlap of CXR acquisitions between the two patient cohorts (see Figure 4.5). The histograms were normalized to have equal area. 84
4.10 UMAP visualization of the confusion matrix for (a) the original test set and (b) the current test set. A similar decision variable was chosen by the deep net for both patient cohorts, classifying positive cases from negative cases (division between blue and orange dots). Overall, the model returned a higher percentage of TPs and lower percentage of FPs for the original test set than for the current test set [127]. 85
4.11 Histogram of patient age from the original dataset (which includes training, validation, and test cases) and current test set [127]. 86

A.1 The comparisons performed between the initial partition of Set A and Set B. The AUC value refers to the test set of Set A, Set A_{te}. The asterisk denotes the statistically significant difference between Set A_{te} and Set B. The green and orange boxes indicate results on Set A_{te} and Set B, respectively. 104
A.2 Summary of the four experiments conducted in this study. 106

A.3 Summary of the results when recalculating (left) and fine-tuning (right) the phase 3 weights of the model. AUC values calculated in the task of distinguishing COVID+/- CXRs were significantly lower when comparing the partitioned Set B (Set $B_{te,I}$ and Set $B_{te,II}$) results with Set A_{te}, denoted by the asterisks. Green and orange boxes indicate results on Set A and Set B, respectively. 108
A.4 Distributions of the AUC values obtained when repartitioning Set $A_{tr,IV}$ 200 times and evaluating it on the test set of Set A, Set A_{te}, and the entirety of Set B. 110
A.5 Summary of the results when implementing $L2$ regularization (left) and repartitioning Set A_{tr} 200 times (right). The AUC value of Set B was significantly lower than Set A_{te} for the $L2$ regularization, denoted by the asterisk. The distributions of the Set A_{te} AUC values and Set B AUC values obtained using the repartitioned Set $A_{tr,IV}$ achieved a significant difference, denoted by the asterisk. Green and orange boxes indicate results on Set A and Set B, respectively. 110

LIST OF TABLES

2.1 Absolute percent difference (± standard deviation) of volume, average DSC, and median DSC at six thresholds. (IQR = interquartile range.) 34
2.2 Absolute percent difference (± standard deviation) of volume, average DSC, and median DSC at six thresholds for the subset analysis. (IQR = interquartile range.) 40
2.3 Summary of the temperatures (which estimate confidence of the model) calculated using the NLL for the four validation sets used in the training process of the model. 42

3.1 Patient demographics categorizing patient sex and age characteristics. 51
3.2 Image acquisition characteristics for the patient cohort analyzed in this study. . 53
3.3 Types of models evaluated in the *BAP1* classification task. 56
3.4 Comparisons of the three best-performing classification models: decision tree, Gaussian process, and support vector. The p-values comparing the differences in AUC values were calculated using the DeLong test, with their corresponding confidence intervals (CIs). Significance levels (α) and widths of the CIs were adjusted for multiple comparisons. None of the comparisons achieved statistical significance after correcting for multiple comparisons using Bonferroni-Holm corrections. 59
3.5 The four texture features most often selected during the 149 LOOCV iterations and the frequency each feature was chosen, i.e., the number of iterations in which a feature was selected. 60
3.6 Model performance using various cross-validation approaches. ROC AUC values in the task of differentiating between *BAP1+* and *BAP1-* patients and 95% CIs for the LOOCV were obtained using 2000 bootstrapped samples. For the 10-fold and 5-fold cross-validation, AUC values were acquired by averaging the AUC values per repeat of the cross-validation approach and 95% CIs were obtained by calculating the 2.5% and 97.5% percentile of the distribution of AUC values. . . 61

4.1 Summary of datasets used, categorized by various factors, including type of units used to acquire the images. 70
4.2 Number of patients categorized by manufacturer and CXR exam type for the original and current test set. 70
4.3 Comparisons of model performance for the three classification algorithms: standard images, soft-tissue images, and both types. The p-values comparing the differences in AUC values were calculated using the DeLong test, with their corresponding confidence intervals (CIs) [142]. Significance levels (α) and widths of the CIs were adjusted based on multiple comparisons. 77
4.4 Performance of the image type-based classification algorithms for the CXR exam types on the current test set. The 95% CIs are displayed in brackets. Majority of portable images were acquired on Canon Inc. units, and majority of DES images were acquired on GE Healthcare and Fujifilm Coporation units. AUC values failed to achieve a statistically significant difference between exam types for each classification algorithm. 78

4.5 AUC values calculated for each of the four variants underlying the two test sets. 79
4.6 Summary of statistical analyses performed between the original test set and the current test set. AUC comparisons were conducted using the unpaired DeLong test, and all three comparisons between the test sets were statistically significantly different ($p < 0.001$). 81
4.7 Definition of PXS score bins for the test sets. 82
4.8 Statistics of the age distributions for the original dataset (which includes training, validation, and test cases) and the current test set. 86
4.9 Distributions of sex for the original and the current test set with their corresponding AUC values and COVID prevalence. 87

A.1 Number of patients and COVID prevalence for Set A and Set B. 102
A.2 Summary of the datasets used and comparisons performed. Of note, cases from the MIDRC Grand Challenge were assessed using the original model, which was pre-trained on the initial partition of Set A, Set A_{tr}. 106
A.3 Summary of the main strategies or applications and their corresponding AUC values. 111

ACKNOWLEDGMENTS

First and foremost I would like to thank my advisor, Dr. Samuel Armato, for the constant guidance and support throughout my Ph.D. career. Sam has provided me with invaluable direction and instruction in conducting this work, and for that, I am very grateful to have been a part of this research and under his tutelage. I would like to extend my thanks to my dissertation committee members, Drs. Maryellen Giger, Hedy Kindler, Patrick La Rivière, and Christopher Straus. I would like to express gratitude to Drs. Kindler and Straus for providing necessary and relevant clinical input and for Dr. La Rivière for his support as my committee chair.

Dr. Giger's contributions and co-mentorship for the third specific aim of this dissertation cannot be understated. Through that work, I was able to fully appreciate the intricacies and nuances of deep learning applications, especially when used in relation to the very prevalent pandemic. The work on the third specific aim also could not have been possible without the members of the Giger lab, specifically Drs. Karen Drukker, Jordan Fuhrman, Hui Li, and Heather Whitney. I was able to gain much insight about the technical and big picture aspects for that work through my many conversations with them.

I would like to thank members of the Armato lab: Linnea Kremer, Roger Engelmann, Dr. Feng Li, and Adam Starkey. The work in the first and second specific aims could not have been completed without the tumor annotations provided by Dr. Li, the software assistance provided by Adam, and the methodology discussions with Roger. I would also like to thank Linnea for the constant support and friendship provided during our shared time as labmates. I want to acknowledge the summer students that I have mentored and that have contributed to this work: Abbas Shaikh, Ilana Deutsch, Isabella Flerlage, and Aditi Kaveti.

I want to recognize the significant contributions made to my professional development as I continue my career in becoming an academic, clinical physicist. Specifically, I would like to thank Drs. Zheng Feng Lu, Ingrid Reiser, Emily Marshall, and Kevin Little. They provided

me with the necessary mentorship during my time here. I am also thankful for the faculty at the Graduate Program in Medical Physics (GPMP) that have helped me in my academic and research education. In addition, I would like to thank Dr. Nicholas Gruszauskas for his assistance regarding any and all image curation procedures at the University of Chicago Medicine. I am also grateful to Owen Mitchell for his data curation efforts and many discussions regarding my second specific aim.

I want to extend my gratitude to my cohort at the GPMP. In particular, I want to thank Natalie Baughan, Julian Bertini, Hadley DeBrosse, Linnea Kremer, and Mira Liu for the friendship, experiences, and encouragement provided during our trying times as we pursued a Ph.D. during a global pandemic. To Mira, I am forever thankful for our enlightening and riveting conversations about any and all things professional and personal. To other classmates I have had the pleasure of interacting with during my time as a GPMP student, I am grateful. I would also like to thank the administrative staff for their continued assistance: Dr. Lili González, Julie Hlavaty, Elena Rizzo, and Chun-Wai Chan.

Lastly, I would like to acknowledge the endless support given to me by my family and friends. To my parents and siblings, I thank them for their routine checkups and providing me with the strength and drive to pursue more in life. I thank my friends for our weekly Zoom meetings, which were initiated due to the unfortunate circumstances of the pandemic, but became a welcome respite, and my roommate who has been a substantial part of this endeavor. To any and all unnamed individuals who have helped on this journey, thank you.

Research reported in this dissertation was made possible by the National Cancer Institute (NCI) of the National Institutes of Health (NIH) under Award Numbers U10CA180821 and U10CA180882 (to the Alliance for Clinical Trials in Oncology), UG1CA189863, UG1CA233320, and UG1CA233253, the John D. Cooney and the firm of Cooney and Conway through the University of Chicago Comprehensive Cancer Center, the NIH S10 OD025081 Shared Instrument Grant, the Hodges Society of the Department of Radiology at the University of

Chicago, and the Lawrence H. Lanzl Graduate Fellowship Award. The work was also part of the Medical Imaging Data Resource Center (MIDRC) and was, in part, made possible by the National Institute of Biomedical Imaging and Bioengineering (NIBIB) of the NIH under contracts 75N92020C00008 and 75N92020C00021. This work has also been supported in part by a C3.AI Digital Transformation Institute award.

ABSTRACT

In this work, we analyze the ability of an automated deep learning-based model to identify the three-dimensional spatial extent of pleural mesothelioma (PM) as presented on computed tomography (CT) scans, employ machine learning to classify an image-based biomarker for PM, and evaluate another model's generalizability in the task of classifying COVID-19 based on patients' chest radiographs (CXRs).

PM is an aggressive form of cancer present in the pleural lining of the lung. It is usually the result of exposure to asbestos and has a very poor prognosis. Linear measurements are the clinical standard used in evaluating tumor response to therapy, but these measurements are only a surrogate for tumor volume. Tumor volume must be calculated to assess tumor burden completely and quantitatively. Determining the volume of tumor, however, is complicated and time consuming, since discerning PM tumor on a medical image is a challenge for human and computer observers alike due to its complex and irregular morphology.

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), an RNA virus that can impact mammals and birds, is the virus responsible for the COVID-19 global pandemic. The primary mode of transmission among humans is through exposure to respiratory fluids carrying infectious virus. Before widespread use of reverse transcription polymerase chain reaction (RT-PCR) tests, CXRs were recommended for triage, disease monitoring, and assessment of concomitant lung abnormalities (e.g., consolidation, ground-glass opacities, and pulmonary nodules).

In recent years, there has been a substantial increase in the application of artificial intelligence and machine learning techniques to medical imaging. Convolutional neural networks (CNNs), specifically, have been successfully employed for various objectives performed on medical images (e.g., the tasks of classification and segmentation). CNNs are capable of learning both local and global patterns of an image, which is essential to identify nuanced disease presentations. In this work, we utilized deep learning methods to assess a novel and

automated pipeline to segment PM. From the segmentations, we obtained tumor volume efficiently and accurately, with the potential of future applications in assessment of tumor extent and response to therapy. Analysis of pixels within the segmented tumor regions was valuable in determining the underlying genetics of the tumor. For example, machine learning techniques and texture analysis applied to the segmented regions determined the mutation status of the BRCA1-associated protein-1 (*BAP1*) gene. The *BAP1* gene is a prognostic factor in PM and can directly impact treatment options for patients. Methods for addressing model generalizability were also assessed for the COVID-19 work, addressing robustness when testing different and newer datasets.

The specific aims of this work were: (1) to implement and study a deep learning model for the automatic segmentation of tumor volumetry in PM, (2) to investigate image texture analysis for differentiation of *BAP1* mutation status, and (3) to evaluate a deep learning model for the generalizable classification of COVID-19. Aim 1 fully investigated the performance of the deep learning model used for PM segmentation, which better informed us of the generated outputs by the model. Aim 2 performed texture feature analysis to determine the somatic *BAP1* mutation status based on the segmented region on a CT scan. Aim 3 evaluated the performance of a deep learning model in the task of COVID-19 classification in order to address and develop methods to achieve model generalizability. These results provided a novel pipeline with potential impact on the future treatment of patients presenting with mesothelioma or COVID-19, expediting many of the sequential steps a patient must undergo and improving the individualized prognostication process, which can also be implemented for other cancers and lung abnormalities.

CHAPTER 1

INTRODUCTION

1.1 Artificial Intelligence

The field of artificial intelligence (AI) originated in the 1950s with early computer scientists exploring the possibility of making computers "think" and perform intellectual tasks traditionally done by humans. AI encompasses various approaches, including machine learning (ML) and deep learning (DL) (Figure 1.1), as well as methods that do not include any learning. Though, for nearly 30 years, AI researchers believed that human-level AI could only be achieved by humans (i.e., programmers) manually creating explicit rules for the program to follow. These rules would mimic human intelligence by implementing logic and if-then rules. For example, early chess programs only involved hardcoded rules designed by programmers, which served as an example of AI but not ML. This paradigm of AI was coined symbolic AI.

Symbolic AI, however, was deemed to be insufficient as it faced limitations in solving complex problems. This included image classification, speech recognition, and language translation (the United States Defense Advance Research Projects Agency [DARPA] initially wanted AI to be used to create autonomous tanks and to automate the translation of Russian to English for intelligence operations [1]). Therefore, these shortcomings led to machine learning.

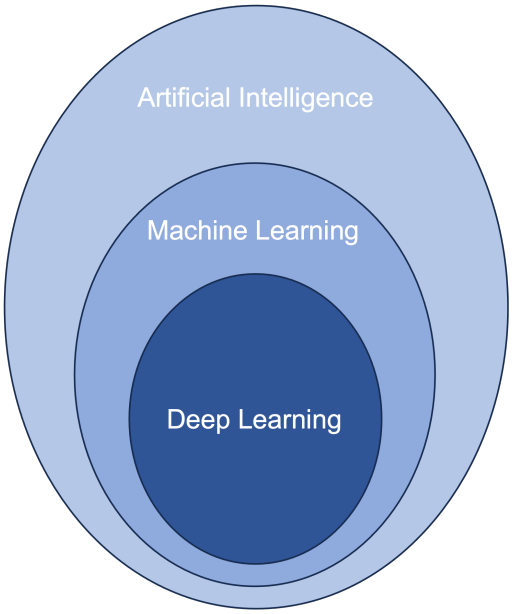


Figure 1.1: Visual representation of AI, ML, and DL, where each field is a subset of the prior.

1.2 Machine Learning

For a machine to go beyond the very explicit rules defined by humans, it must be able to learn the patterns itself to perform a specified task. For instance, in the symbolic AI model, the rules and data were input to the machine, and answers were output. In ML, the data and answers are provided to the machine, and the rules are then "learned," hence, the machine learning, or training, on the input data. The rules are created through statistical measures, relating the given data with the answers provided.

To assess the performance of the ML algorithm, a loss function is introduced to measure the distance between the predicted output of the algorithm to the expected output, or "truth." This distance can then be used to adjust the algorithm (i.e., "weights") to improve the results and minimize the distance. This process is the "learning" of machine learning. We can appreciate this by visualizing the linear least squares method, a common loss function,

as shown in Figure 1.2, which assumes the following linear system:

$$\vec{y} = X\vec{w}. \tag{1.1}$$

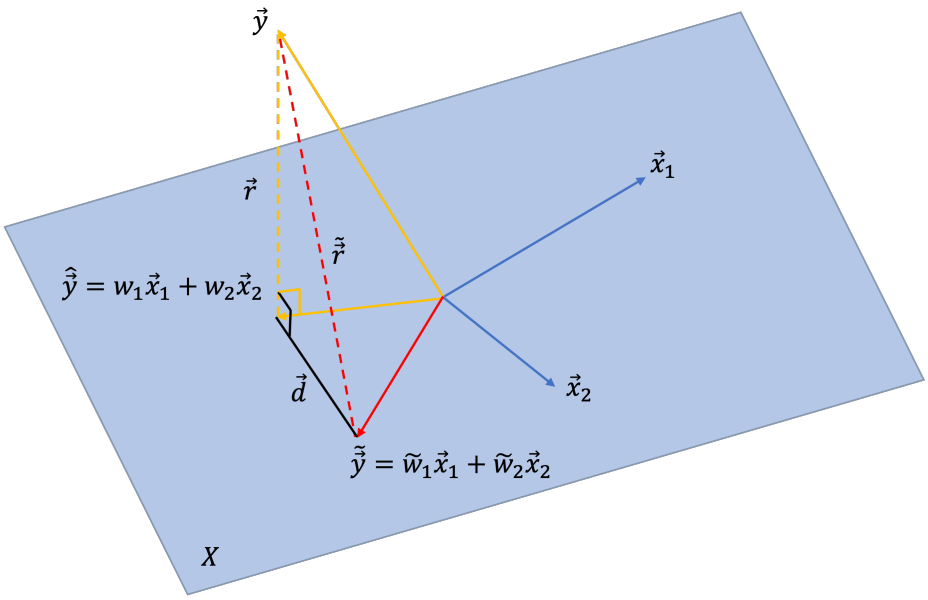


Figure 1.2: Geometry of least squares. The schema displays the columns of dataset X, which span the blue hyperplane. The truth label vector is projected on this hyperplane. The distance between the weights resulting from the projection ($\vec{w}$) and nonoptimal weights ($\tilde{\vec{w}}$) due to a residual is also shown.

As shown in Figure 1.2, $\vec{y} \in \mathbb{R}^n$ is the truth label vector, $X \in \mathbb{R}^{n \times p}$ is the matrix of features, $\vec{w} \in \mathbb{R}^p$ is the weights vector, and $\vec{r}$ is the residual vector defined as $\vec{r}_i = \vec{r}_i(\vec{w}) = y_i - \langle \vec{w}_i, \vec{x}_i \rangle$, where $\vec{r}_i$ is the residual for i-th equation of n linear equations. For simplicity, the span of X is reduced to two columns ($p = 2$), where $\vec{x}_1$ and $\vec{x}_2$ are the columns of the subspace X. The vector $\hat{\vec{y}}$ is the orthogonal projection of $\vec{y}$ onto the subspace.

Predictions $\tilde{\vec{y}}$ with residual $\tilde{\vec{r}}$ depict nonoptimal weights as $\tilde{\vec{r}}$ is not perpendicular to X. This can be shown by using the Pythagorean theorem: $||\tilde{\vec{r}}||^2 = ||\vec{r}||^2 + ||\vec{d}||^2 \rightarrow ||\tilde{\vec{r}}|| > ||\vec{r}||$ as $\vec{d} > 0$, which indicates that the weights resulting in $\tilde{\vec{y}}$ cannot be optimal as the distance between $\vec{y}$ and hyperplane X is not minimized and $||\tilde{\vec{r}}||^2$ is not as small as possible (i.e.,

there is no right angle with the hyperplane). Therefore, we aim to solve for the weights that minimize the following residual:

$$||\vec{r}||^2 = \sum_{i}^{n} |r_i(w_i)|^2. \tag{1.2}$$

This results in this unique, closed form solution for $\text{rank}(X) = p$ (i.e., X has p linearly independent columns):

$$\hat{\vec{w}} = \arg\min_{\vec{w}} ||\vec{y} - X\vec{w}||_2^2 \tag{1.3a}$$

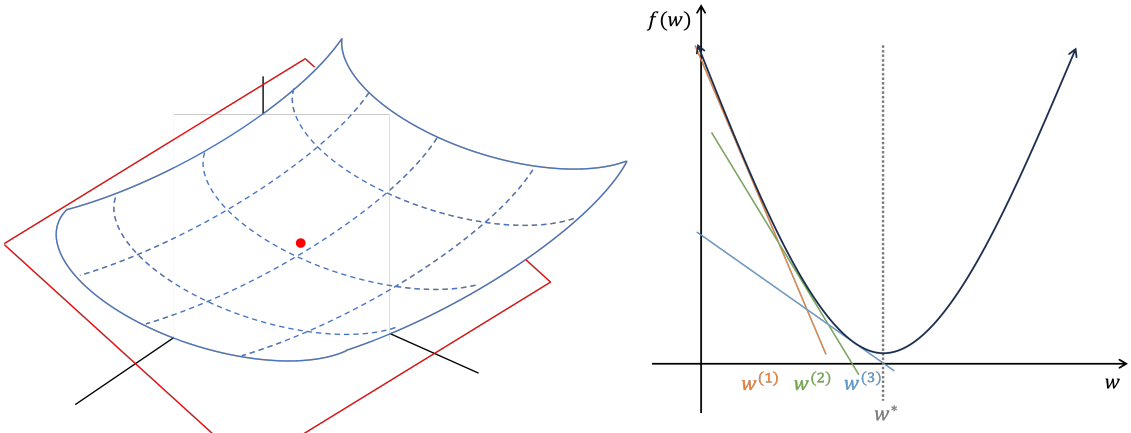
$$= (X^\top X)^{-1} X^\top \vec{y}. \tag{1.3b}$$

Therefore, the least squares example demonstrated how machine learning can be used to train and learn the best weights given the input data and the answers. This example can also be extended to implementing gradient descent, which is a method that can easily perform convex optimization, or minimizing convex functions (Figure 1.3). Gradient descent is used to analytically find the combination of weight values that yields the smallest possible loss function [2]. For least squares, gradient descent is advantageous because it avoids calculating the inverse of matrices. This can be shown by defining the convex function $f(\vec{w}) = ||\vec{y} - X\vec{w}||_2^2$ (Equation 1.3a) with solution $\vec{w}^* = (X^\top X^{-1})^\top \vec{y}$ (Equation 1.3b). If $f(\vec{w})$ is expanded as:

$$f(\vec{w}) = \vec{y}^\top \vec{y} - 2\vec{w}^\top X^\top \vec{y} + \vec{w}^\top X^\top X\vec{w}, \tag{1.4}$$

then its gradient is easily shown to be:

$$\nabla_{\vec{w}} f = 0 - 2X^\top \vec{y} + 2X^\top X\vec{w}. \tag{1.5}$$



(a) Red dot in the center shows the intersection of the convex function with the red hyperplane.

(b) Gradient descent starts with initial guess $w^{(1)}$ and repeatedly takes steps in the direction of the negative gradient.

Figure 1.3: Visualization of gradient descent with (a) displaying a 3D convex function that intersects a hyperplane, where the intersection shows the lowest value the loss function can take. The gradient descent process can be visualized in 2D in (b), with w^* showing the lowest value of the convex loss function $f(\vec{w})$.

Gradient descent is also favorable because of its generalizability as it is one of the most used optimization algorithms for ML, and specifically, DL, with stochastic gradient descent (SGD) often employed [3].

This least squares example also demonstrates how ML algorithms transform input data to achieve the desired outputs. For example, recall the hyperplane in Figure 1.2 was the subspace spanned by the columns of X, with $\hat{\vec{y}}$ as the orthogonal projection of $\vec{y}$ onto the subspace. This displays how higher-dimensional data can have different, and often lower-dimensional, representations. In other words, ML algorithms attempt to meaningfully transform the data to achieve improved outputs closer to the truth. One popular method for dimensionality reduction, i.e., representing the data into a lower dimensional subspace, is principal component analysis (PCA). PCA creates an orthogonal projection, or transformation, of data such that the variance of the projected data is maximized along the principal components. PCA, and dimensionality reduction in general, is an example of unsupervised learning in ML. Unsupervised learning is the branch of ML that attempts to learn "inter-

esting" properties of an underlying distribution of the data at hand. More formally, an unsupervised algorithm observes several examples of random vector $\vec{\boldsymbol{x}}$ and aims to learn the probability distribution $p(\vec{\boldsymbol{x}})$ from which $\vec{\boldsymbol{x}}$ arose. In contrast, supervised learning also observes several examples of random vector $\vec{\boldsymbol{x}}$, but takes into consideration the truth (or label) vector $\vec{y}$, and learns to map the two. Therefore, the algorithm is able to predict $\vec{y}$ given $\vec{\boldsymbol{x}}$, $p(\vec{y}|\vec{\boldsymbol{x}})$ [3]. Solving for the optimal weights in the least squares example is a display of supervised learning.

1.2.1 Cross-validation

To validate how well a trained model performs on an unseen dataset (i.e., model generalizability) and to adjust any additional hyperparameters of the model, cross-validation (CV) can be performed. In CV, which is a resampling procedure, the dataset is first split between a training set and a validation (or test) set. The model then gets trained on the training set and separately evaluated on the validation set. There are many different forms of CV. The most common are: hold-out, k-fold, leave-one-out, stratified k-fold, and nested k-folds. Nested k-fold CV, in particular, is done to adjust hyperparameters of the model that may need tuning (e.g., number of estimators used by a model or number of features to select). A visual representation of k-fold CV is shown in Figure 1.4. To address class imbalance, stratified k-fold ensures that the same percentage of samples of each target class as the complete set exists per fold. Lastly, k-fold CV and stratified k-fold CV can be performed iteratively, where a random portion of the dataset is split for each k-fold, and that is performed for a certain number of predefined iterations: this is called repeated k-fold. It is advised to perform hold-out CV when there is a large dataset size, k-fold when there are few samples for the validation to be reliable, and repeated k-fold to ensure robust performance of the model when there is a small dataset size [2].

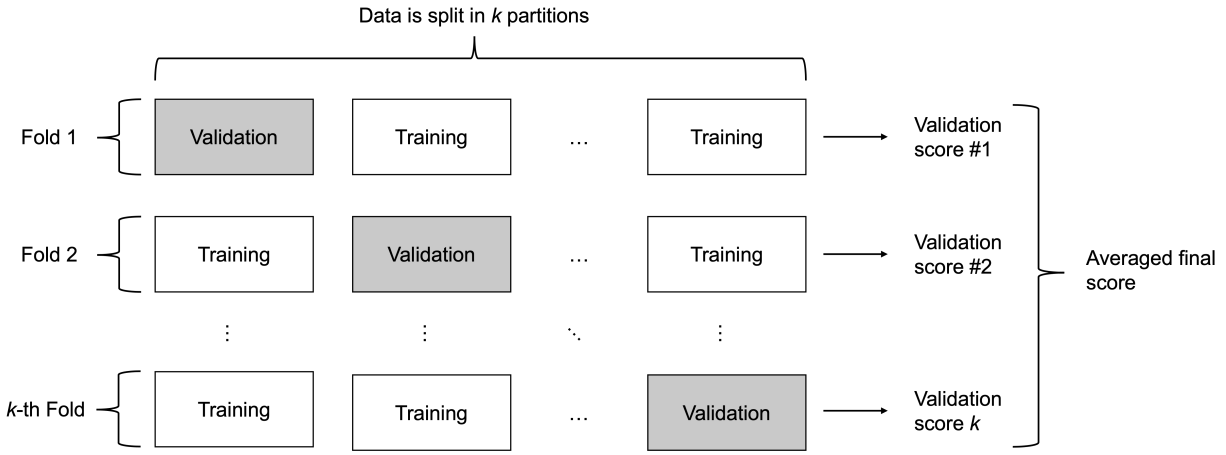


Figure 1.4: Schema depicting k-fold CV. For leave-one-out CV, k would equal the number of samples n, and $n-1$ samples would be used for training, with the n-th sample used for testing.

1.3 Deep Learning

The least squares example highlighted in Section 1.2 explained three things: 1) the idea and importance of a loss function, 2) the versatility of gradient descent, and 3) the ability to transform data into representations more interpretable by the model. It is imperative to appreciate these same three concepts when discussing DL. Expounding on the third point, the "deep" of deep learning simply refers to the number of successive layers that DL algorithms utilize when training. Conceptually, these layers create and learn different representations of the input data, with the number of layers on the order of tens to hundreds.

In DL, these layers learn the representations of the data through models called neural networks. Neural networks consist of multiple layers of "nodes," starting with an input layer, hidden layers in the middle, and lastly, an output layer. Each node consists of a weight $\vec{w}$ multiplied with the input $\vec{x}^\top$ and summed with a "bias" factor b, which is all placed in an

activation function σ. This output (y in Equation 1.6) becomes the input of the next layer:

$$y = \sigma(\vec{x}^\top \vec{w} + b). \tag{1.6}$$

A simple neural network is shown in Figure 1.5.

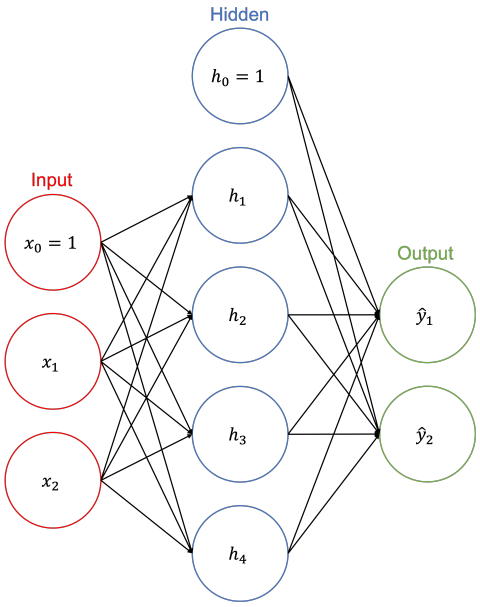


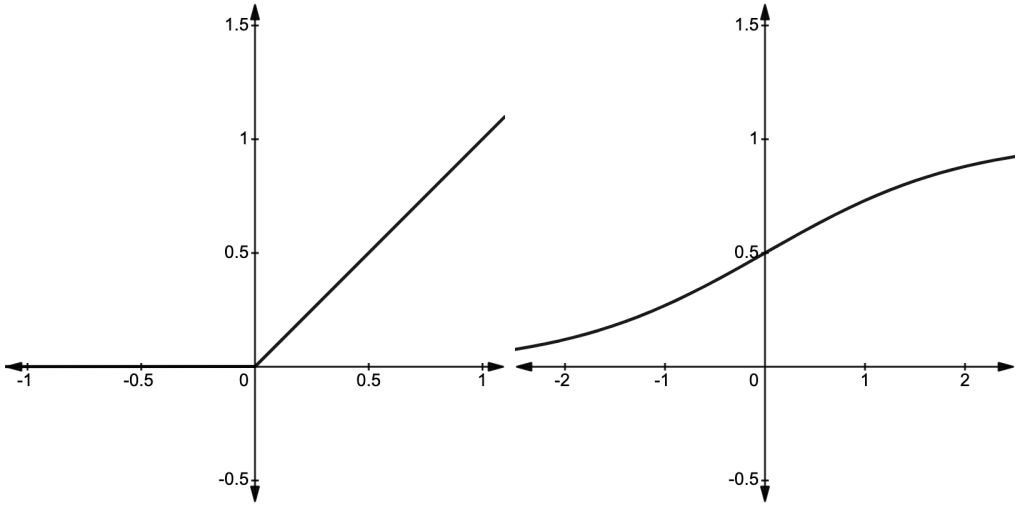
Figure 1.5: An example of a neural network with one hidden layer in blue. The values $x_0 = 1$ and $h_0 = 1$ are the bias terms used in the model. Of note, this figure depicts "fully connected layers" as every input neuron is connected to every output neuron.

Without activation functions, models can only learn and represent the data in a linear pattern, i.e., the dot product between the input and weights with an addition of the bias, as shown in Equation 1.6. Therefore, activation functions are necessary as they expand the ability of the layers in the model to represent the data in a nonlinear fashion. The most common activation function in DL is the rectified linear unit (ReLU), which is used to zero out negative values (see Figure 1.6a and Equation 1.7):

$$y = \max(0, x). \tag{1.7}$$

The sigmoid activation function, often used for binary classification tasks (as it pertains to the work presented in this dissertation), places arbitrary values into a [0, 1] interval (see Figure 1.6b and Equation 1.8), outputting numbers that may be interpreted as a probability (see Section 1.3.2):

$$y = \frac{1}{1 + e^{-x}}. \tag{1.8}$$



(a) ReLU activation function. (b) Sigmoid activation function.

Figure 1.6: Common nonlinear activation functions used in DL algorithms.

To learn the optimal weights, DL models first undergo the "forward pass." That is, the neural network first makes predictions using randomly initialized weights and a set value for the bias (typically 0) [4]. Then, the initialized weights and biases are evaluated using a loss function (see Section 1.2). For binary classification, the network will attempt to minimize the binary cross-entropy L:

$$L(y_i, p_i) = -(y_i \log(p_i) + (1 - y_i) \log(1 - p_i)), \tag{1.9}$$

where y_i is the binary indicator (0 or 1) denoting the class for the sample and p_i is the probability of the target class for that sample. In particular, SGD can be used to minimize

the loss, similar to the gradient descent process of least squares. The SGD here is called the optimizer. The gradients of the SGD are calculated using the backpropagation algorithm. Collectively, the weights and biases of the model are constantly adjusted and updated to minimize differences between predictions and truth.

1.3.1 Convolutional neural networks

Convolutions are powerful in computer vision tasks as they can capture spatial information and local patterns presented in images. This is in contrast to Dense layers (Figure 1.5), which only learn global patterns as inputs are reduced to flattened vectors. Because of their ability to capture spatial information, convolutions 1) are translation invariant and 2) can capture spatial hierarchies of patterns. In other words, once a pattern is learned by the network, it will always recognize it. Further, spatial hierarchies are recognized through the consecutive layers, with small local patterns such as edges being learned in the first layers, and larger patterns made of features from the earlier layers are captured by deeper layers. Overall, a convolutional neural network (CNN) is constructed of four types of layers: convolution, activation function (Figure 1.6), pooling, and fully-connected (Figure 1.5).

Convolution filters are applied to images initially input to the network. These images (also known as input feature maps) can be defined as 3D tensors, having dimensions of *height* by *width* by *depth*. For RGB images, the *depth* corresponds to the three color channels. Medical images only have one channel, by definition, as they are grayscale. A number of convolutions can be applied per image, therefore, the output feature maps will no longer have depth that corresponds to the number of color channels. Rather, the depth is arbitrary and will depend on the number of convolution filters applied. These filters will become the weights of the network and will learn patterns present in the image such as edges, or more complex shapes in deeper layers, during backpropagation.

The stride of the filter refers to how many pixels the filter slides over when applied to the input feature maps. For example, a stride of 1 means the filters are applied in a pixel-wise manner, as shown in Figure 1.7. A filter with stride 2 will be applied every two pixels. The larger the stride, the smaller the output feature map becomes. (It is uncommon, however, to have strides larger than 3 [5].) The size of the output feature map can be controlled by zero-padding the input map, i.e., by adding a certain amount of rows and columns of zeros around the edges of the input. Overall, the relationship between the input feature map size (W), convolution filter size (F), stride (S), amount of zero-padding (P), and the output feature map size can be represented as: $\frac{W-F+2P}{S} + 1$. Between the sequential convolution and activation function layers, pooling layers are typically applied on the feature maps to reduce the dimensionality of the representations learned by the filters. Pooling layers are usually 2×2 and can be either done using max (downsampling the feature maps by selecting the largest value in the 2×2 pools) or average (downsampling the feature maps by selecting the average value in the 2×2 pools) pooling. In contrast, "unpooling" can be performed to upsample the feature maps to output an image of desired size.

CNNs have been used for image classification and segmentation tasks. Though, segmentation can be thought of as pixel-wise classification (Figure 1.8). For semantic segmentation, CNNs assign a prediction value per pixel for each class label in the image. For instance segmentation, models are trained to assign prediction values for objects of the same class. These predictions are made on the logit vectors, or the "raw" outputs, of the DL algorithm before being input to a final activation function. The final activation functions are typically sigmoid (Equation 1.8) or softmax. Interpretation of these prediction values is discussed more in Section 1.3.2.

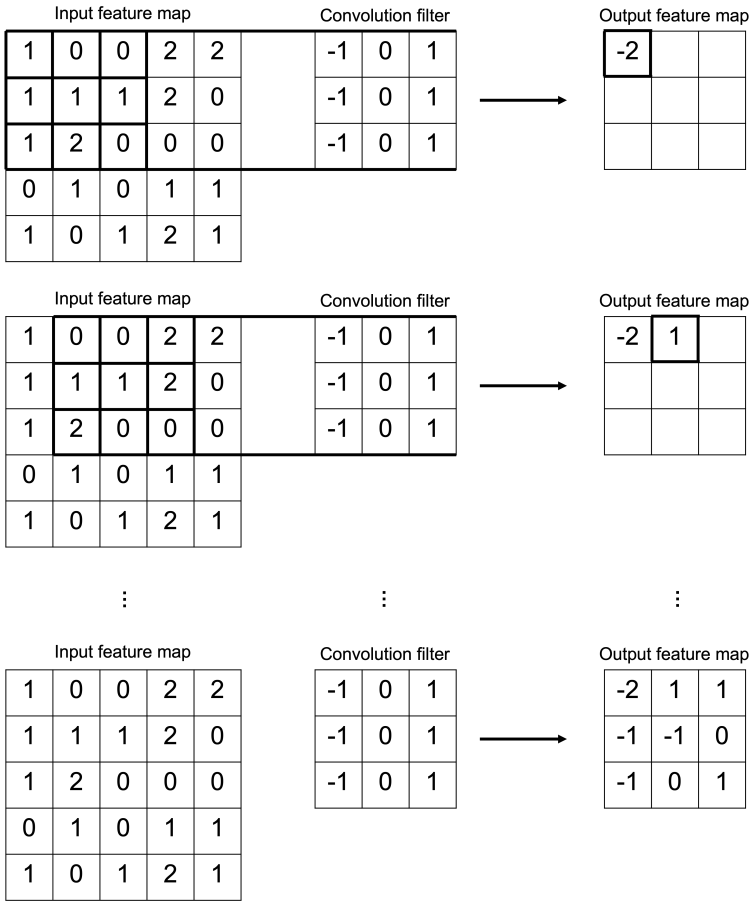


Figure 1.7: Demonstration of a 2D vertical edge detection convolution filter operating on an input feature map. The filter was applied with stride 1, as the second row in the figure displays the movement across one pixel when compared to the top row. Further, no zero-padding was applied to the input feature map, which resulted in a "valid" convolution. A valid convolution is defined as a convolution only performed over pixels where the convolution filter overlaps completely with the input feature map—values outside the filter have no effect on the output feature map. The bottom row of the figure displays the output feature map after the convolution filter is applied to the entirety of the input feature map.

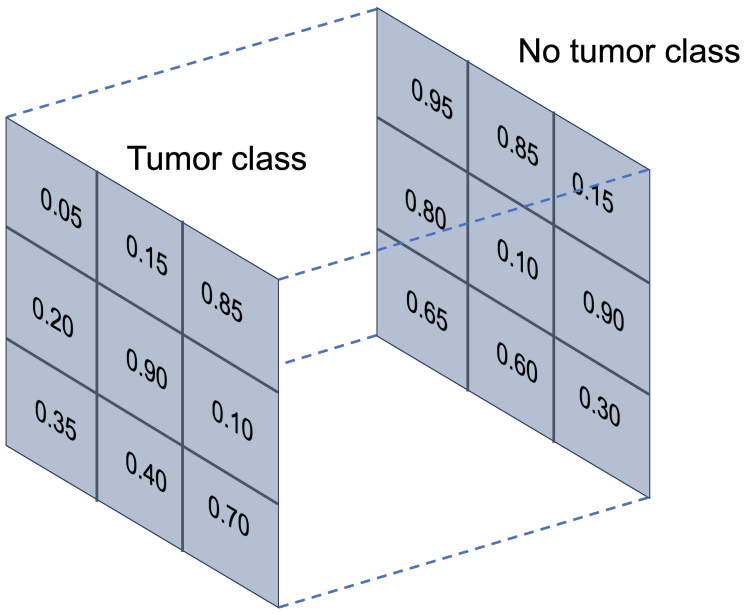


Figure 1.8: Probability maps for two potential classes ("tumor" and "no tumor") in medical image semantic segmentation tasks. These values would be the result of the logits input to the final layer, i.e., activation function, of the model. Because this is a binary task, the pixel-wise values across the two class channels will add to 1. The argmax function could be used to return the final binary segmentation.

1.3.2 Model calibration

Model (or confidence) calibration addresses "the problem of predicting probability estimates [that are not] representative of the true correctness likelihood" [6]. In other words, the probability associated with the predicted class label should reflect its true correctness likelihood [6, 7]. Calibration is also important as the confidence estimates returned by the models can be used for model interpretability. For instance, deep CNNs may be used by an autonomous driving system to analyze real-time scenes captured by cameras [8]: the semantic segmentation performed of street scenes should yield accurate detection of pedestrians and other vehicles. It is equally crucial for the system to discern instances where these predictions may lack reliability. Another example is the segmentation of brain tumors also using CNNs [9]. If the CNN struggles to confidently delineate critical areas of the brain, it becomes

imperative for a medical expert to intervene or be notified about these uncertain regions. Therefore, semantic segmentation networks need to generate both accurate label predictions and reliable confidence measures.

A common challenge for modern CNNs designed for semantic segmentation is that the models often exhibit overconfidence in their predicted labels, which is due to overfitting [6, 10, 11, 12]. However, classification models have exhibited the same behavior. Therefore, many of the methods used for model calibration in classification tasks have translated over for segmentation, the most popular being temperature scaling (TS), a post-hoc processing technique [6, 13, 14]. Post-hoc processing is one methodology among many, such as regularized training and Bayesian modeling [15]. A main difference between these techniques is when the calibration is performed. For example, regularized training has calibration done during training whereas post-hoc techniques have it performed during the validation phase, after training, but prior to testing.

TS is an extension of Platt scaling [16], a method used for calibrating ML algorithms, not solely deep nets. Platt scaling is a parametric approach to calibration: nonprobabilistic (i.e., uncalibrated) predictions of a classifier are used as features for a logistic regression model, which is trained on the validation set to return probabilities. Specifically, Platt scaling learns scalar parameters $a, b \in \mathbb{R}$ and outputs $\hat{q}_i = \sigma(az_i + b)$, where σ is the sigmoid activation function, $z_i \in \mathbb{R}$ is the model's non-probabilistic output (known as the *logit* vector), and $\hat{q}_i$ is the calibrated probability. Parameters a and b can be optimized during a loss function, typically the negative log-likelihood (NLL) loss, over the validation set. Of note, the learned model weights are frozen during this stage as this calibration method is post-hoc.

Extending Platt scaling results in TS as it only uses a single scalar parameter T, the temperature, where $T > 0$ for all classes. Given the logit vector $\boldsymbol{z}_i$ for sample i, class k, and

the softmax activation function σ_{SM}, the new confidence prediction is [6]:

$$\hat{q}_i = \max_k \sigma_{\mathrm{SM}}(\boldsymbol{z}_i/T)^{(k)}. \tag{1.10}$$

The calculated temperature "softens" [6, 17] the output of the last activation layer with $T > 1$ reducing model confidence (i.e., probability predictions), $T = 1$ indicating no change to the original probability, and $T < 1$ increasing model confidence. Overall, as $T \to \infty$, the probability $\hat{q}_i$ approaches $1/K$, which indicates maximum uncertainty, and as $T \to 0$, the probability collapses to a "point mass" ($\hat{q}_i = 1$) [6]. As with Platt scaling, T is optimized with respect to the NLL on the validation set. Importantly, since T does not change the maximum of the softmax function (or sigmoid, as the two activation functions are equal for binary cases), the calibrated class prediction for sample i, $\hat{y}_i^{'}$, is not impacted. In other words, TS does not affect the model's accuracy.

Translating TS for a segmentation task results in the following formulation [7]:

$$T^* = \arg\min_T \left(-\sum_{i=1}^{n} \sum_{x \in \Omega} \log \left(\sigma_{SM}(\boldsymbol{z}_i(x)/T)^{(S_i(x))} \right) \right), \tag{1.11}$$

where the optimal T is once more calculated by minimizing the NLL described in Equation 1.11 with respect to a hold-out validation set. The variable Ω denotes the image space, n the number of training images, x is location, and $S_i(x)$ is the true predicted segmentation ("truth" image) for image i at location x where $x \in \Omega$. In this definition, however, temperature scaling assumes that each image has the same temperature. Therefore, there are many different and more advanced temperature scaling approaches, such as local TS (where $T_i(x) \in \mathbb{R}^+$ is image and location dependent) [7], entropy-based TS (a method that scales the confidence of a prediction according to its entropy) [14], selective TS (which introduces a binary classifier as a selector to categorize correct and incorrect predictions for separate scaling) [15], and attended TS (which works properly for small-size validation sets, highly accurate deep CNNs,

and validation sets with noisy labels) [13]. The latter three methods are currently only used for classification tasks.

1.4 Texture Feature Analysis

Medical images are highly quantitative mathematical constructs that allow for a range of computer science and biomedical engineering investigations. The underlying numeric data associated with pixels in an image may be explored through a variety of statistical measures and quantitative features, which are collectively called radiomics. There are two methods to produce quantitative features: the conventional method or with deep learning. Conventional features—often related to the texture of an image—imply standard statistics of pixel values such as the average gray level intensity. Conversely, deep learning quantitative features are acquired from the convolutional layers of the deep network. This discussion will focus on conventional methods.

While texture may be perceived qualitatively [18], it was initially quantified with predefined rules certain algorithms follow (similar to rule-based algorithms used for chess, as mentioned in Section 1.1) [19]. An early example of this rule-based implementation was used to determine malignancy status in mammograms [20]. For example, a manual cutoff threshold may be applied to a feature (e.g., spiculation), and if the feature value from a region of interest (ROI) is lower than the threshold, the ROI is deemed benign [19, 20]. Radiologic appearance of a tissue (e.g., dysplasia of breast parenchyma) also has been used to construct a quantitative measure that could be used as a predictive marker for risk of malignant tumor [21, 22, 23].

More recently, capturing texture has evolved into quantified features that are generated using mathematical formulations. Some of these features are based on the gray level histogram (first-order), gray level co-occurrence matrix (GLCM), fractal analysis, Laws' texture energy measures, and power law spectrum [24, 25, 26, 27]. Overall, these features attempt

to capture the coarseness, consistency, and arrangement of pixels within the image [28]. In practice, the pipeline for texture feature extraction from a medical image is as follows: image acquisition, image preprocessing, drawing an ROI, feature extraction, feature analysis/classification, computer output [23]. Using feature values for classification follows the same fundamental machine learning methods as discussed in Section 1.2. For instance, in supervised learning, the extracted feature values are input to an ML algorithm with the corresponding truth labels, and the algorithm will use the learned "rules" with which it classifies the data. Common ML algorithms used for feature selection and classification in medical image texture feature analysis include linear discriminant analysis (LDA), stepwise linear regression, k-nearest neighbor, artificial neural networks (ANNs), and support vector machines (SVMs) [19, 24, 29, 30].

As with other ML tasks, it is important to take model overfitting into consideration [19, 31]. Since texture features can number in the hundreds [32] and thousands, which may exceed the number of samples, the models may only optimize to the values of features presented, reducing model generalizability (i.e., ability of the model to perform strongly on an independent test set). This phenomenon is called the "curse of dimensionality." Overall, texture features have been extracted from many different imaging modalities, including chest radiographs (CXRs), mammograms, ultrasound, and computed tomography (CT) scans. Texture features have also been used to evaluate many different diseases. The usage of texture analysis and/or AI on medical images to inform clinicians of pertinent information is called computer-aided diagnosis (CADx).

1.5 The Role of AI in Medicine

AI has played a substantial role in helping radiologists in the context of diagnosis (CADx), abnormality detection (CADe), triaging (CADt), and acquisition and optimization of images (CADa/o) since the mid-1980s [23, 33]. While AI was first implemented to analyze lung

and breast cancers, its reach has since been extended to include an array of diseases and abnormalities, such as diabetic retinopathy, polycystic kidney disease, prostate cancer, and head-and-neck cancers [34, 35]. With advancements in computer technology and processing power, CNNs, the leading DL technique, were first introduced to medical imaging in the mid-1990s, successfully identifying lung nodules on chest radiographs and microcalcifications on mammograms [36, 37]. CNNs are desirable as they do not require hand-crafted features to be calculated and recorded; CNNs learn from the inputs and truth labels provided to them during training (Section 1.3.1), with the expectation that their performance will generalize to novel cases. For older CADx systems, the quantitative rules were manually designed, and image analysis was conducted. In contrast, by providing a CNN with the input data and the expected output, the CNN constructs its own rules that help to transform the data into meaningful results, hence the learning, as discussed in Sections 1.2 and 1.3. CNNs have been applied to images acquired from different modalities and of different anatomic regions, performing segmentation and classification tasks [38, 39, 40, 41]. For example, previous studies have displayed the successful segmentation of pleural mesothelioma (PM) tumors as displayed on CT scans [42, 43]. CNNs have also been involved with the COVID-19 pandemic, providing diagnoses of the disease based on patients' CXRs [44]. Accordingly, the application of CNNs is an essential preliminary step in automated segmentation of PM, which can be used to quickly assess tumor burden and evaluate a patient's response to treatment. In addition, with the abundance of medical images acquired during the pandemic, model generalizability and robustness can be evaluated in the task of COVID-19 diagnosis.

1.6 Clinical Pipeline, Characterization, and Potential for AI in the Assessment of Mesothelioma

Patients with PM typically first present with dyspnea and/or chest pain [45, 46]. Initial diagnosis may involve visual assessment of CXRs, where the image often reveals a unilateral

pleural effusion [46, 47, 48]. Diagnosis for PM is confirmed by a tumor biopsy through thoracoscopy [46, 49]. The biopsy also classifies the tumor histologic subtype (e.g., epithelioid, sarcomatoid, or biphasic), as that is the most reliable prognostic factor for mesothelioma [46, 49]. There is ongoing research toward biomarker evaluation in PM. For example, the BRCA1-associated protein-1 (BAP1) is a deubiquitinase, controlling cell growth, cell proliferation, and cell death. The *BAP1* gene is of great interest in the field of mesothelioma since it is the most mutated somatic gene in PM. *BAP1* mutations can also be inherited, and individuals with germline mutations in this gene have been widely recognized as being predisposed to the disease and other cancers; though, studies have suggested that germline mutations of *BAP1* are associated with better prognosis for patients with mesothelioma [50, 51].

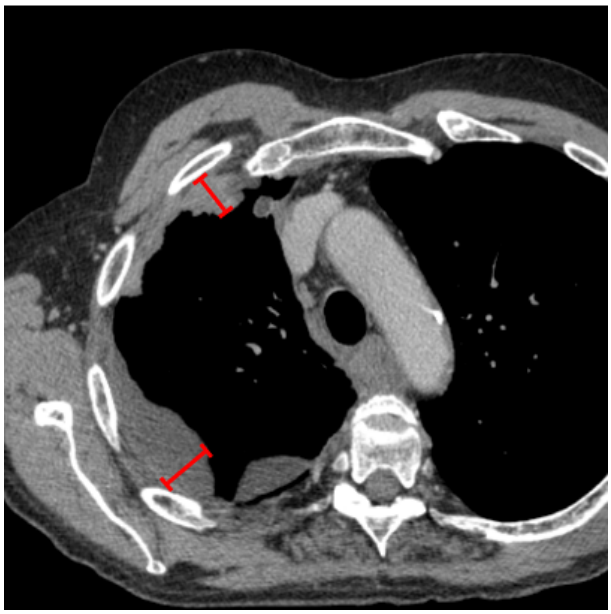


Figure 1.9: Linear measurements made by a radiologist to quantify tumor burden. Using longitudinal summations of these measurements, patient response is evaluated and the efficacy of treatment is assessed.

Recent clinical guidelines strongly recommend that initial staging be obtained using a CT scan [51], as a CT scan is more sensitive and specific for imaging mesothelioma than

a CXR [52]. Imaging studies are also essential when evaluating patients longitudinally, as the efficacy of treatment can be determined by change in tumor burden as presented on the images. In current clinical practice, tumor burden is captured through linear measurements of PM, and the linear measurements are recorded on CT sections to evaluate tumor burden (Figure 1.9). These measurements are a surrogate for tumor volume, which would offer a more complete assessment of disease burden. Tumor volume could be calculated through either manual or semi-automated analysis by a radiologist. For instance, the radiologist tasked with manual analysis must segment the tumor, through manually delineating the tumor boundary, throughout an entire CT scan (which may comprise over 200 sections). For semi-automated analysis, the radiologist provides initial input to a computer algorithm or modifies output produced by the computer or both.

Manual or semi-automated delineation of PM on CT scans, however, is an arduous task. First, the morphology and presentation of mesothelioma tumor is irregular and difficult to outline. Its appearance is also challenging to discern as there is low contrast between the tumor and adjacent soft tissue and pleural effusion. Second, an observer must consistently delineate the tumor across all CT scan timepoints to reliably measure change in tumor burden, which leads to an appropriate assessment of response to therapy. Collectively, this approach is too time-consuming and burdensome for use in the clinic. These difficulties can be mitigated using DL, specifically CNNs. This is a crucial step in automating tumor volume measurements as the laborious part is performed automatically, and simple pixel counting remains for tumor volume calculations. Further, identification of the mutation status through image analysis can help aid decision-making with regard to genetic testing, which in turn can create a more customized plan for treatment and the preemptive assessment of family members.

1.7 Clinical Pipeline, Characterization, and Potential for AI in the Assessment of COVID-19

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), an RNA virus that can impact mammals and birds, is the virus responsible for the COVID-19 global pandemic. The primary mode of transmission among humans is through exposure to respiratory fluids carrying infectious virus. The virus is highly contagious and can rapidly mutate. Further, infection with the virus has led to severe and/or fatal disease. Early detection of the disease can mitigate the symptoms, however, and patient prognosis can improve. Before the widespread use of reverse transcription polymerase chain reaction (RT-PCR) tests, CXRs were recommended for triage, disease monitoring, and assessment of concomitant lung abnormalities (e.g., consolidation, ground-glass opacities, and pulmonary nodules) [53, 54]. In addition, CXRs are widely accessible, which makes them an ideal modality for an image-based evaluation of the disease.

The pandemic has resulted in the acquisition of many medical images: this diversity of images allows for a comprehensive evaluation of AI models as we are able to assess the generalizability of the models when utilizing the various datasets available. The rapid abundance of datasets, however, introduces new challenges for data curation and truth labeling. Further, the various patient demographics and different imaging parameters should be incorporated when developing AI systems. Therefore, while the need for early and reliable diagnosis of COVID-19 has been met with the use of RT-PCRs, DL can still be used for the detection of COVID-19-related lung abnormalities as presented on CXRs. The images also can be used to create a robust AI model, and a pipeline of data curation and model evaluation can be established.

1.8 Outline of Dissertation

This dissertation discusses the implementation and performance of machine and deep learning techniques in medical image analysis, as it pertains to CT scans of mesothelioma and CXRs of COVID-19.

Chapter 2 investigates the usage of CNNs for the automated segmentation and volume calculation of PM tumor as presented on CT scans. To evaluate the performance of the DL model, two figures of merit are employed: percent difference of volume and the Dice similarity coefficient. Using these two metrics, the segmentations are compared with a reference standard provided by an experienced radiologist. The segmentations produced by the CNN are binarized at various probability thresholds to assess the impact of the thresholds. The confidence of the CNN is also evaluated to determine whether the output probabilities are properly calibrated. The implementation of a robust deep CNN for the segmentation of PM tumor should result in a rapid calculation of tumor volume, which can improve patient outcome as volume has been shown to be a more accurate metric for assessing tumor burden and response.

Chapter 3 presents a novel methodology for the classification of somatic *BAP1* mutation based on texture features extracted from CT scans of PM patients. While the germline *BAP1* mutation has more clinical impact as patients with germline mutation have improved survival, their family members have a 50% of inheriting the mutation, and the germline mutation status may guide treatment decisions, this proof of concept work displays the potential of image-based assessment of mutation status of the *BAP1* gene. One study has shown the potential of imaging genomics for somatic *BAP1* mutation classification, but the study presented in this chapter is the first to employ DL and ML algorithms in tandem, using the former for the tumor segmentations and the latter for the classification task. The translation of this research to germline *BAP1* mutations has the potential to improve patient prognostication and family member assessment.

Chapter 4 assesses the performance of a separate CNN in the task of COVID-19 classification based on patient CXRs. The model had initial strong performance in predicting COVID-19 status on the original dataset on which it trained. Though, when a larger, and more current dataset from the same institution that provided the original data was tested, performance significantly decreased. To investigate this discrepancy, several factors are considered to compare the two sets of data. These factors include patient demographics, clinical factors, image acquisition dates, and quantifying model perception of the CXRs. This work substantially contributes to the discussion of model robustness and generalizability as the in-depth investigations provide invaluable insight on model performance.

Chapter 5 summarizes the main conclusions and potential future directions of the work presented in this dissertation. As the common thread throughout this dissertation has been the employment and evaluation of ML and DL algorithms for various medical imaging tasks, possible additional research may be performed to improve the methods, especially in terms of model generalizability and robustness to improve care for patients presenting with mesothelioma and other lung abnormalities.

CHAPTER 2

CONVOLUTIONAL NEURAL NETWORKS FOR SEGMENTATION OF PLEURAL MESOTHELIOMA: ANALYSIS OF PROBABILITY MAP THRESHOLDS

2.1 Introduction

Pleural mesothelioma (PM) is an aggressive form of cancer present in the pleural lining of the lungs. It is often the result of exposure to asbestos and has a poor prognosis [55]. Computed tomography (CT) is the most common imaging modality used to stage and assess patients with PM [48, 56, 57, 58]. The current standard to evaluate tumor response to therapy is the modified Response Evaluation Criteria in Solid Tumors (mRECIST) [59, 60]. Using this protocol, clinicians obtain up to six measurements of "tumor thickness residing perpendicular to the chest wall or mediastinum" as presented on a CT scan [59]. These guidelines were more recently updated to mRECIST 1.1 [61] to better align with RECIST 1.1 [62].

In contrast to the linear measurements of mRECIST, manual volumetric analysis conducted by radiologists can be used to better estimate tumor burden and can also be used to obtain image-based biomarkers [63]. Further, tumor volume has displayed strong predictive power in patient assessment in terms of overall and progression-free survival [64, 65]; however, acquisition of manual tumor volume is too time-consuming and burdensome to be systematically used in routine clinical care [66].

Machine learning, and specifically deep learning, techniques have been implemented for various medical image classification and segmentation tasks [38, 39, 40, 41]. Deep learning has been used specifically for tasks in the PM setting, such as improvement of subsampled magnetic resonance image quality, histologic subtype classification on hematoxylin and eosin-stained slides, and prognosis based on 3D positron emission tomography-CT images and clinical data [67, 68, 69]. Similarly, deep learning can be used to mitigate the difficulties

of tumor volume calculations, in particular, using convolutional neural networks (CNNs). Previous studies have successfully implemented CNNs in PM segmentation, which is a crucial step in automating tumor volume measurements as the laborious part is performed automatically, and simple pixel counting remains for the volume calculation [42, 43, 70].

For a CNN to properly segment PM tumor on CT sections, the network must be trained and validated using a labeled set of images. Due to the fundamental statistical nature of machine learning, the outputs of a CNN are probabilities. For instance, in identifying the location of tumor on a CT scan, the CNN assigns each pixel a probability of being tumor. Therefore, for each CT section input to the network, a probability map is generated that displays the likelihood of tumor in a pixel-wise fashion. In practice, a threshold is set to binarize these maps so that any pixel with probability equal to or greater than 0.5, for example, is set to 1 ("tumor"), and all other pixels are set to 0 ("not tumor"). Given that modern neural networks tend to be overconfident, and the output probabilities may not be true estimates of the confidence of a model [6, 7, 71], the 0.5 probability might not generate the most accurate segmentation for a disease as complicated as PM tumor. Therefore, the purpose of the present study was (1) to better understand the probability values returned by the CNN and (2) to investigate whether different probability thresholds could improve pixel-wise class labeling in this complicated tumor. The tumor segmentations obtained using different thresholds were evaluated against a reference standard using the Dice similarity coefficient (DSC) and the percent difference of volume as the figures of merit. Therefore, the impact that varying thresholds would have on the predicted tumor segmentation was investigated, given the inherent complexities of this tumor. Further, preliminary work was performed to assess the confidence of the model using standard temperature scaling (TS) techniques.

Overall, the choice of threshold may have considerable impact on the final tumor segmentation. In terms of PM tumor volumetry, a lower threshold applied to the probability maps

would result in a larger volume as more pixels are now considered tumor; however, this may result in an increase of pixels erroneously labeled as tumor, thus negatively impacting the CNN accuracy of tumor segmentation. In contrast, increasing the threshold may produce a segmentation that is too restrictive, substantially decreasing the tumor volume calculated. Furthermore, the change in threshold alters the overlap of tumor identified by the CNN with the reference standard (as determined by an experienced radiologist). Therefore, this study[1] investigated the impact of probability thresholds on tumor volume and the overlap of tumor contours by applying a broad range of thresholds, recording the volumes and the DSC, and studying the resulting trends.

2.2 Methods

2.2.1 Patient population

The patient cohort was compiled from a previous study performed by the Cancer and Leukemia Group B (CALGB 30901) [73]. CALGB is now part of the Alliance for Clinical Trials in Oncology. The CALGB 30901 study evaluated 49 patients with confirmed unresectable epithelioid, sarcomatoid, or mixed-type PM, and patients were without disease progression after 4 to 6 cycles on first-line therapy with pemetrexed and cisplatin or carboplatin. Patients were randomly assigned to either the treatment arm (continued therapy with pemetrexed alone) or the observation arm. The patients underwent CT scans at baseline and then every 6 weeks for the first 6 months. Each participant signed an Institutional Review Board (IRB)-approved, protocol-specific informed consent document in accordance with federal and institutional guidelines.

The present study was retrospectively conducted on 186 baseline and follow-up CT scans of 48 patients from the CALGB 30901 study. There was an average of 123 sections per scan

1. This chapter is based on a study reported in [72].

(range: 39-696 sections). The most common section thickness was 5 mm (range: 0.625-5 mm). The scans had been acquired using 21 different scanner models.

CNN-derived contours

The CNN used in this study employed the U-Net deep CNN (2D) architecture. Specifically, the deep CNN architecture consisted of a downsampling and upsampling path. For the downsampling path, a Visual Geometry Group 16 (VGG16) model was pre-trained on the ImageNet database using scale-jittering [74, 75]. Layers of the downsampling path were initialized using the weights acquired from the VGG16 training scheme. A 2×2 max pooling operation with stride 2 was applied to the feature maps at each step of the downsampling path. A dropout layer of probability 0.5 was used to prevent overfitting. During the upsampling path, a 2D operation using nearest-neighbor interpolation was applied to the feature maps. The network output a segmentation mask the same size as the input image size (512×512 pixels). A rectified linear unit (ReLU) activation function was applied after each convolutional layer, except for the last layer, which used a sigmoid activation function that returned pixel-wise probabilities on the range [0,1] for the segmentation task, i.e., whether a pixel contains tumor. A threshold value of 0.5 was applied to the output of the sigmoid layer during the validation step so that any pixel with a probability 0.5 or greater was labeled "tumor." During its training phase, the network minimized the binary cross-entropy, which was averaged in a pixel-wise manner between each predicted segmentation and the provided reference standard. Adam, an algorithm for first-order gradient-based stochastic optimization, was used to optimize the network during training using a learning rate of 10^{-5}.

The VGG16/U-Net deep CNN architecture had been previously trained on a completely separate set of 126 PM patients, some presenting with pleural effusion [43]. In this earlier study, the CNN was tested on 77 patients, some of whom presented with both tumor and pleural effusion and some only with tumor; the median DSC and median average Hausdorff

distance for that method were 0.690 and 5.1 mm, respectively, as previously reported [43]. More information about this original model can be found in Gudmundsson et al. [43]. For the present study, tumor contours of the external CALGB dataset were automatically generated and evaluated with no additional training or validation of the model.

Radiologist reference contours

A research radiologist was presented with the initial CNN contours (generated using a probability threshold value of 0.5, the conventional threshold for binary classification and segmentation tasks) and was able to modify or redraw the contours using in-house software to provide the reference standard. Due to the time-consuming nature of adjusting the contours, however, the radiologist was presented with sections separated by approximately 5 mm. This process resulted in an average of 52 reviewed sections per scan (range: 32-70 sections). Contour comparisons and tumor volumes were performed only on sections that the radiologist reviewed.

2.2.2 Model calibration

TS is a post-hoc probability calibration method used for multi-class classification. For medical image semantic segmentation tasks, the two classes would be "disease" or "no disease," as mentioned in Section 2.1. TS estimates a single scalar parameter temperature $T > 0$ using the logit $\boldsymbol{z}_i$ vector as input, where i is the i-th image. The temperature is typically optimized only on the validation images and using the negative log-likelihood (NLL) cost function, as was performed in this work.

The temperature was calculated for the four separate validation sets used to develop the original VGG16/U-Net deep CNN: left or right hemithorax displaying either tumor only or tumor plus effusion. For the left hemithorax, 275 sections displayed tumor only, and 97

sections displayed tumor plus effusion. For the right hemithorax, 216 sections displayed tumor only, and 101 sections displayed tumor plus effusion.

2.2.3 Tumor volume and Dice similarity coefficient

Tumor volume calculation

Tumor volume was defined as:

$$\text{Volume } [\text{mm}^3] = \sum \text{Number of pixels within a contour} \times \text{pixel dimension}^2\,[\text{mm}^2] \times \text{inter-section distance [mm]}, \quad (2.1)$$

where the summation was over all sections containing a contour. The number of pixels within a contour was equal to the number of nonzero pixels within the binary mask created after applying a threshold to the probability maps generated by the CNN. Pixel dimension (in units of mm^2) was acquired from the Digital Imaging and Communications in Medicine (DICOM) image header. Inter-section distance corresponded to the difference in table position between two sections on which the radiologist provided reference contours. Tumor volumes were computed using the CNN-derived contours and the radiologist reference contours. The absolute percent difference of volume was calculated by taking the absolute value of the difference between the reference and the CNN-derived volumes divided by the average of the two. All tumor volume computations were performed using MATLAB (Mathworks Inc., Natick, Massachusetts).

Dice similarity coefficient

Another metric used to compare the CNN tumor contours generated at the different thresholds with the radiologist's reference standard was the DSC [76]. The DSC was calculated for

each individual CT section (using MATLAB's "dice" function), and a final DSC was calculated per patient scan after averaging the DSC values across all sections. Figure 2.1 displays the overall pipeline conducted in this study.

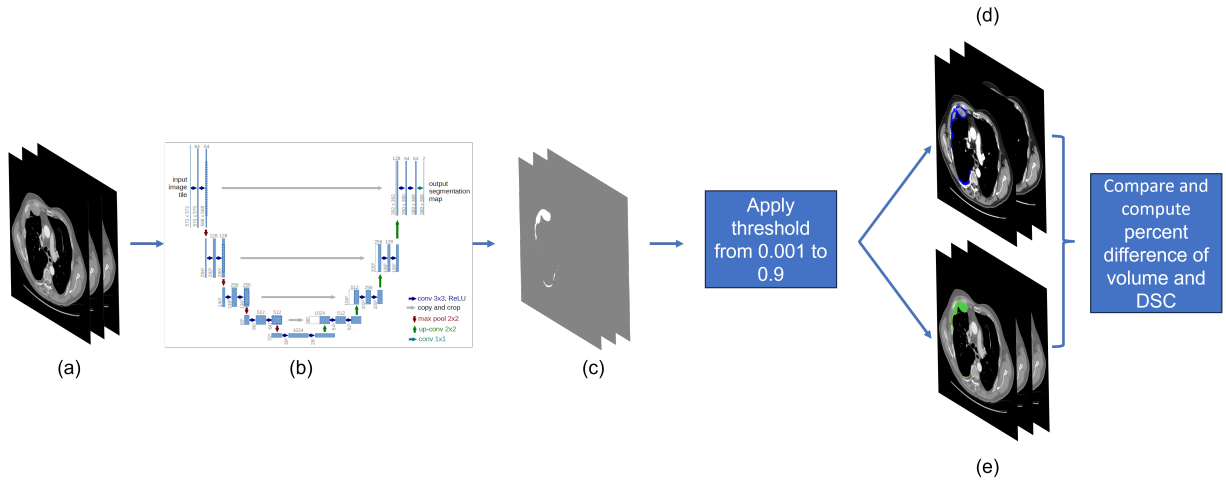


Figure 2.1: Schema demonstrating the methodology employed. Beginning from left to right: (a) 2D CT sections of a patient were input to the (b) VGG16/U-Net, and (c) the probability maps were generated. The probability maps were binarized using a range of thresholds, where (d) the reference standard was provided by a radiologist by modifying the generated segmentations at the 0.5 threshold. Lastly, the reference standard was compared to (e) the probability maps binarized at the various thresholds, using the percent difference of volume and DSC as the two figures of merit. U-Net figure reprinted, with permission, from [74].

2.2.4 Statistical methods

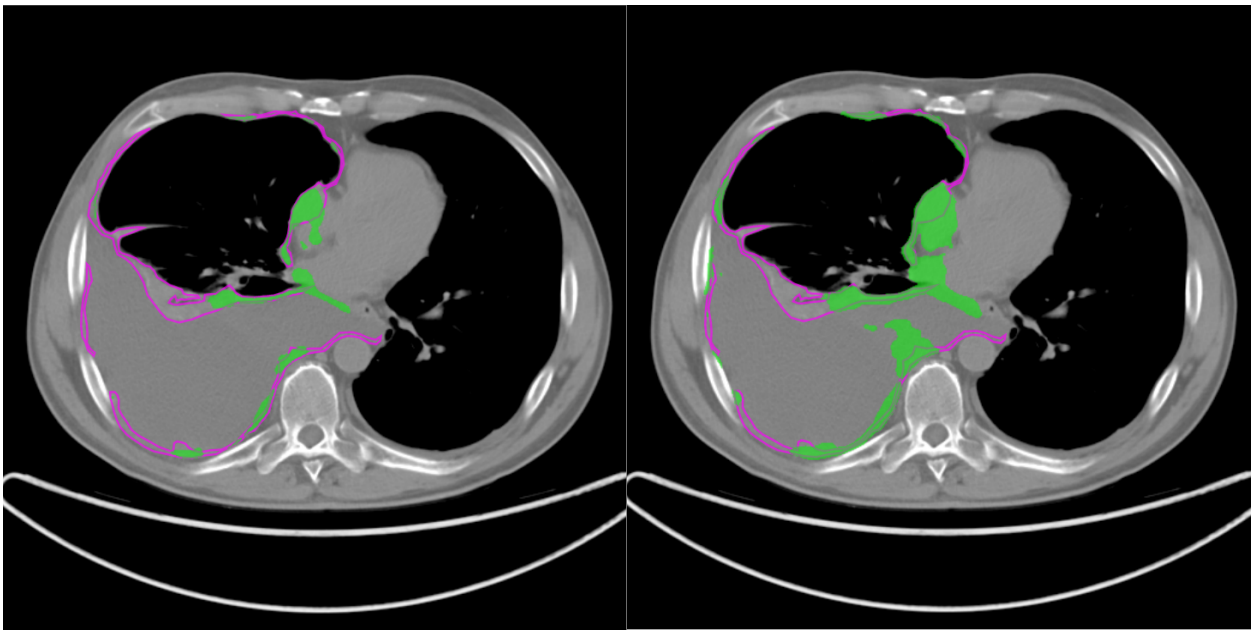
Comparison of the absolute percent difference of volumes and comparison of DSC values across thresholds were first checked for normality using the one-sample Kolmogorov-Smirnov test [77]. Since the null hypothesis was rejected, the data did not come from a standard normal distribution. Therefore, the Wilcoxon signed-rank test was used to compare DSC or absolute percent volume differences computed between the reference standard contours and contours generated across a range of CNN probability map threshold values. The significance of p-values was adjusted using the Bonferroni correction to account for 15 comparisons, and

statistical significance was considered at $p = 0.0033$. Data collection was conducted by the Alliance Statistics and Data Management Center. Data quality was ensured by review of data by the Alliance Statistics and Data Management Center and by the study chairperson following Alliance policies.

2.3 Results

2.3.1 Tumor volume and DSC

Figure 2.2 displays a visual representation of a change in the probability thresholds and its impact on the tumor contour generated by the CNN. Pleural effusions present were difficult for the CNN to fully capture as shown. Overall, the thresholds ranged from 0.001 to 0.9; however, the CNN never assigned any pixel a probability of 0.75 or greater. Figure 2.3 shows boxplots of the DSC values comparing the reference contours to contours obtained from the CNN-generated probability maps at six thresholds. Except for the 0.01 threshold, the range of DSC values decreased with the incremental reduction of probability thresholds. The median did not substantially change (see Table 2.1). In particular, the DSC values at the 0.1 threshold did not achieve a significant difference with threshold values other than 0.01, as shown in Figure 2.4b.



(a) CNN tumor segmentation at probability threshold of 0.5.

(b) CNN tumor segmentation at probability threshold of 0.001.

Figure 2.2: Differing contours on the same section of the same patient created with an adjustment in the CNN probability threshold. Purple represents the radiologist reference outline, and green represents the CNN pixel-wise segmentation prediction of tumor with (a) a probability threshold of 0.5 (average DSC over all sections: 0.357) and (b) a probability threshold of 0.001 (average DSC over all sections: 0.476).

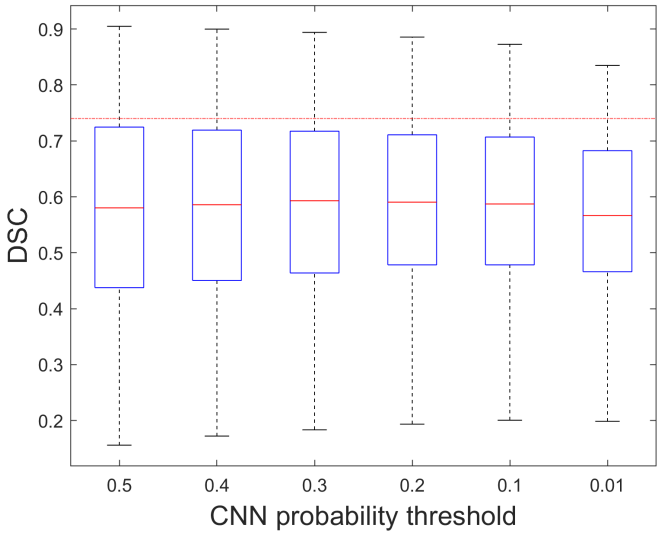


Figure 2.3: Boxplots showing the DSC values obtained for tumor comparisons acquired between the radiologist and the deep CNN at six different thresholds. The solid red lines display the median DSC value at each probability threshold. The dashed red line displays an average human interobserver DSC of 0.74 achieved between radiologists in the task of segmenting mesothelioma on CT scans from a separate dataset [42].

Table 2.1 shows that the average absolute percent difference of volume consistently decreased with a lower threshold, demonstrating the underestimation of the tumor by the CNN at the conventional 0.5 threshold. The underestimation by the CNN is also demonstrated in Figure 2.5. While there was strong linear correlation as determined by the Pearson correlation coefficient ($r = 0.89$, $p < 0.001$), the volumes calculated from the radiologist reference contours were larger than the volumes obtained from the CNN-derived contours at the 0.5 threshold. There was one outlier (not shown), which was a case presenting with severe disease. The underestimation was not as prominent at the 0.01 threshold (Figure 2.4a). The average DSC peaked at the 0.2 threshold, while the median DSC reached its maximum at the 0.3 threshold. Figure 2.4 displays all the relevant p-values for pairwise comparisons of the percent difference of volume and DSC values. Both metrics were affected by changes in threshold, though at different threshold comparisons.

Table 2.1: Absolute percent difference (± standard deviation) of volume, average DSC, and median DSC at six thresholds. (IQR = interquartile range.)

	Threshold					
	0.5	0.4	0.3	0.2	0.1	0.01
Average absolute % difference of volume	42.93 ± 32.99	36.75 ± 30.88	31.18 ± 28.17	26.01 ± 24.67	22.09 ± 19.11	26.60 ± 17.17
Average DSC	0.58 ± 0.17	0.59 ± 0.17	0.59 ± 0.16	0.59 ± 0.16	0.59 ± 0.15	0.56 ± 0.14
Median DSC (IQR)	0.58 (0.29)	0.59 (0.27)	0.59 (0.25)	0.59 (0.23)	0.59 (0.23)	0.57 (0.22)

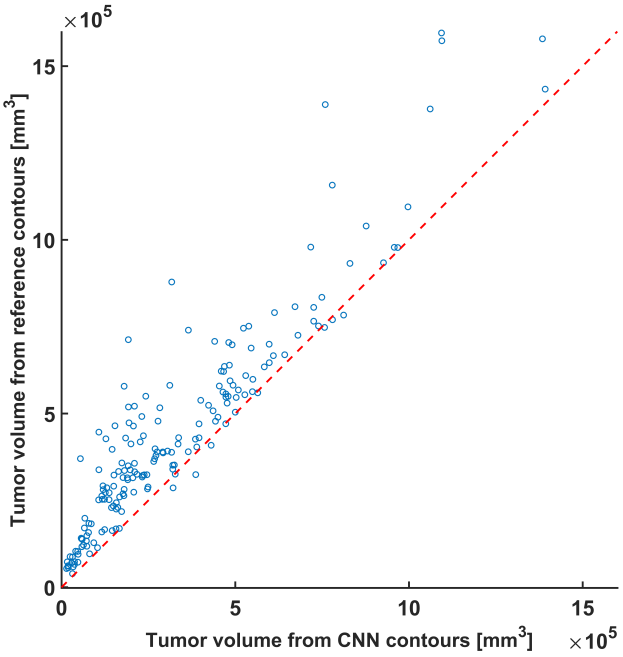
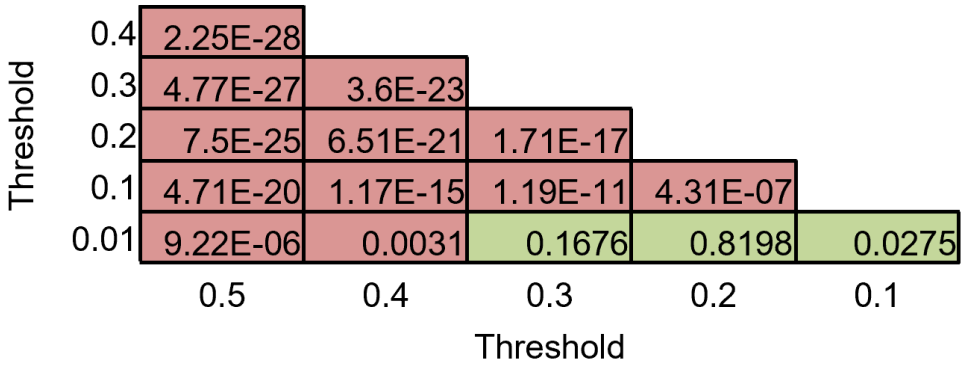
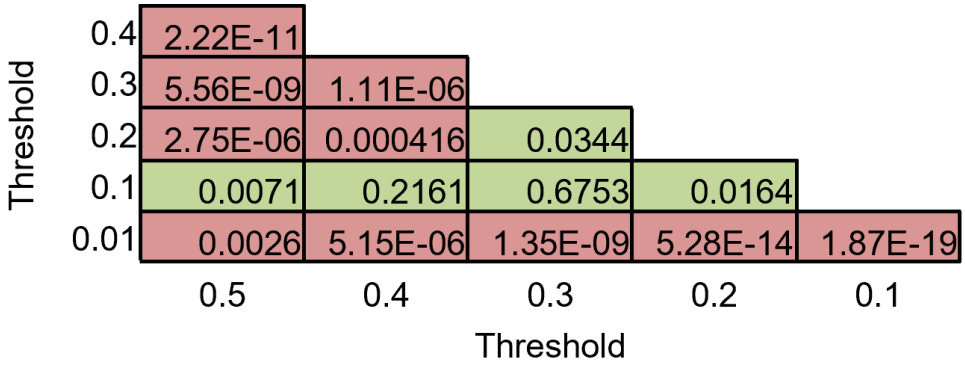


Figure 2.5: Scatter plot displaying the correlation between the tumor volumes calculated from the CNN contours obtained at the 0.5 threshold and the radiologist reference contours. One outlier at (14.2×10^5 mm^3, 33.1 $\times10^5$ mm^3) is not shown. The dashed red line represents the identity line.



(a) P-values for the absolute percent difference of volume.



(b) P-values for the DSC values.

Figure 2.4: Matrix of p-values when comparing the absolute percent difference of volume (a) and DSC (b) across thresholds. Red indicates a significant difference ($p < 0.0033$ after Bonferroni correction), and green indicates a failure to achieve significance, as determined by the Wilcoxon signed-rank test.

Figure 2.6 is a Bland-Altman plot [78] displaying the relative percent difference of volume as calculated using the radiologist and the CNN contours at the 0.5 threshold. The CNN consistently underestimated the tumor at this threshold, which resulted in a mean percent difference of 42.5% (range from -17.5% to 148.8%, median of 34.76%). Seventeen (9.14%) scans were within ±5% of 0% difference as shown by the red band in Figure 2.6.

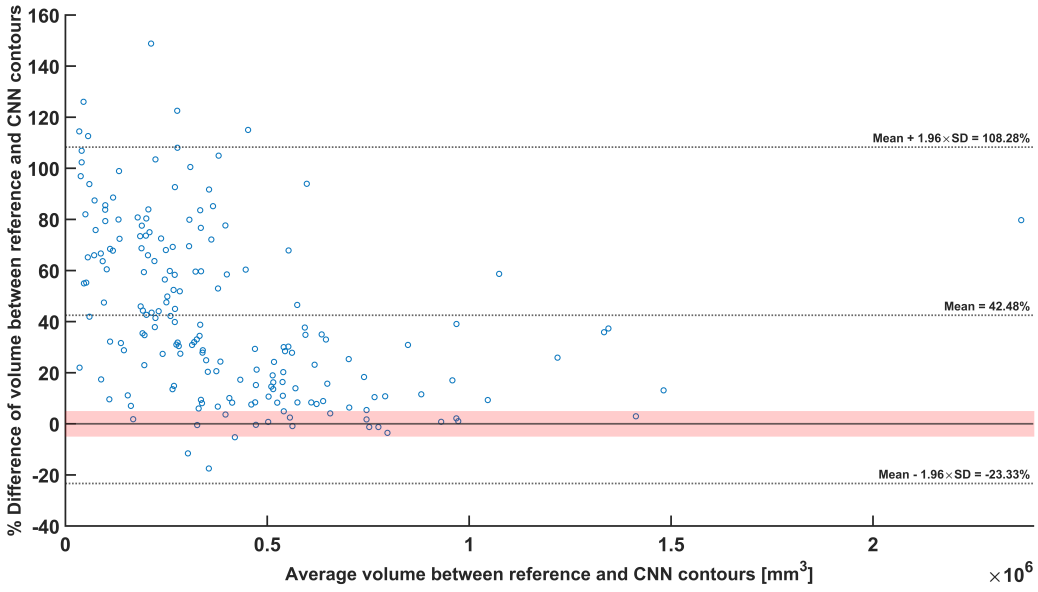


Figure 2.6: Bland-Altman plot of the relative differences between reference and CNN-based tumor volumes at the 0.5 threshold. The red band highlights differences within 5% of 0.

Figure 2.7 shows the frequency of a threshold being chosen as "optimal," ranging from 0.001 to 0.7. An optimal threshold differed based on the metric. For example, a certain threshold being optimal for volume meant this was the threshold that resulted in the lowest percent difference of tumor volume between the radiologist and CNN contours; similarly, an optimal threshold for DSC indicated that it maximized DSC. Beside the substantial peak present at a threshold of 0.5 for the DSC, there did not appear to be a distinct pattern of "best" thresholds. When inspecting the average of each metric across all cases, maximum

DSC occurred at a 0.2 threshold, while the lowest absolute percent difference occurred at a 0.06 threshold (Figure 2.8).

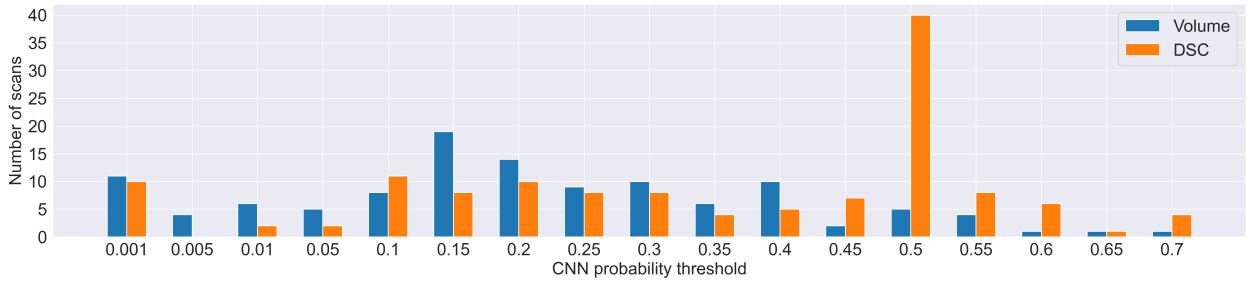


Figure 2.7: Histogram of the CNN output thresholds that maximize DSC and minimize percent difference of volume.

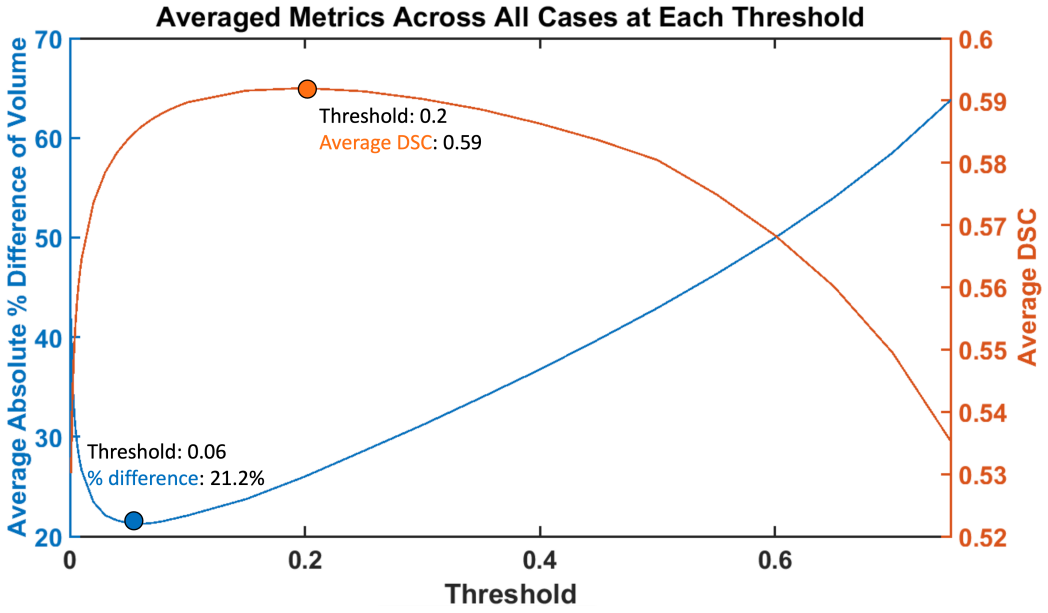


Figure 2.8: The average absolute percent difference of volume (and its minimum) and the average DSC (and its maximum) across all cases for the entire threshold range.

Figure 2.9 displays example images from four of the six scans (from four patients) that were greater than the 95% agreement limits as presented in the Bland-Altman plot in Figure 2.6. The CNN was trained on cases that were applicable to mRECIST measurements and

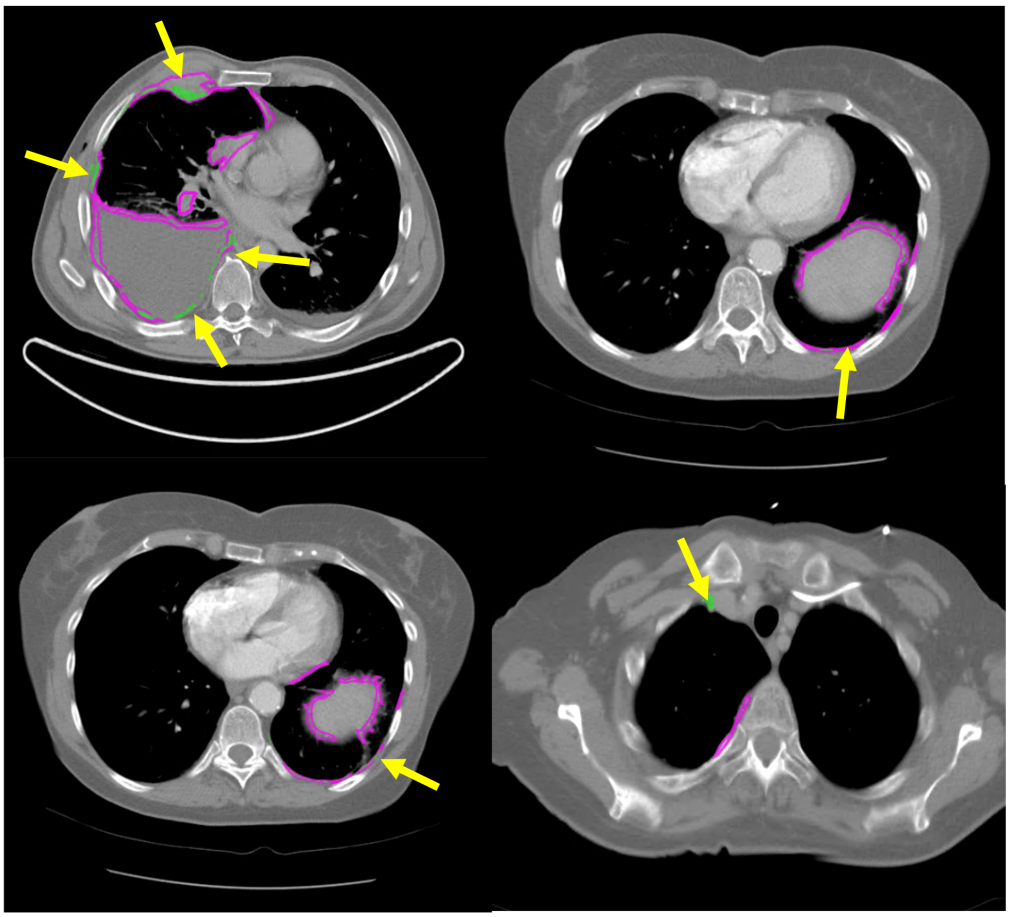


Figure 2.9: Example images from four of the six scans that exceeded the 95% agreement limits as shown in the Bland-Altman plot (Figure 2.6). Yellow arrows point to regions where the CNN predicted tumor at the 0.5 threshold. Purple outlines are the radiologist reference contours.

was not designed to consider tumors present near the diaphragm, near the lung apices, or invading other parts of thoracic anatomy. Therefore, the CNN may have been confounded by cases with severe pleural effusion as well as disease superior to the aortic arch and inferior to the pulmonary vein, where other anatomic structures complicate the morphology of PM. To exclude such regions, the volume and DSC comparisons only considering CT sections inferior to the aortic arch and superior to the pulmonary vein were performed. Figures 2.10-2.11 and Table 2.2 parallel Figures 2.3-2.4 and Table 2.1, showing DSC and volume values for the subset analysis. Figure 2.10 displays the DSC across the same six thresholds analyzed

in Figure 2.3, and Table 2.2 displays the average absolute percent difference of volume along with average and median DSC values across the thresholds. While the range of DSC values was slightly larger for the subset analysis (range from 0.17–0.93 versus 0.16–0.90), the highest overall DSC achieved was for the subset analysis at a DSC of 0.93. DSC values were slightly more robust for the subset analysis across thresholds, as fewer of the comparisons in Figure 2.4b resulted in statistically significant differences compared with those in Figure 2.11b. All significant pairwise comparisons of percent volume differences were the same between Figures 2.4a and 2.11a.

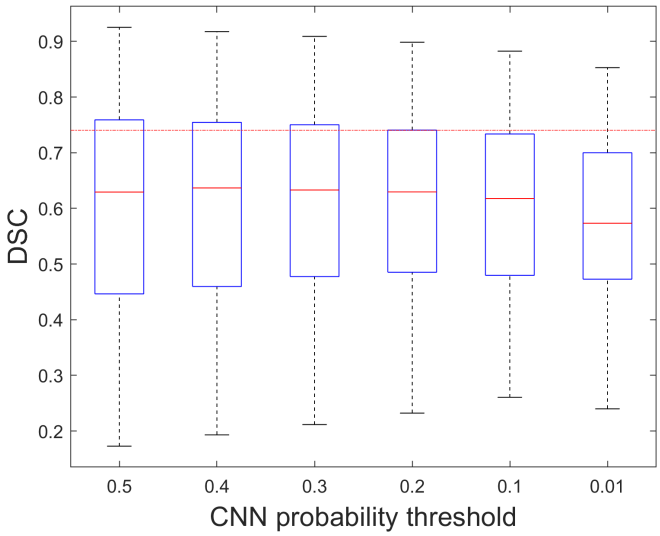
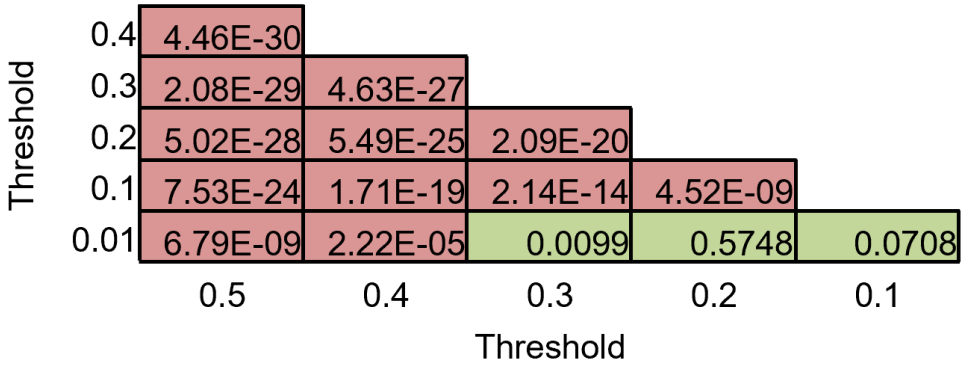


Figure 2.10: Boxplots showing the DSC values obtained for the subset tumor comparisons acquired between the radiologist and the deep CNN at six different thresholds. The solid red lines display the median DSC value at each probability threshold. The dashed red line displays an average human interobserver DSC of 0.74 achieved between radiologists in the task of segmenting mesothelioma on CT scans from a separate dataset [42].

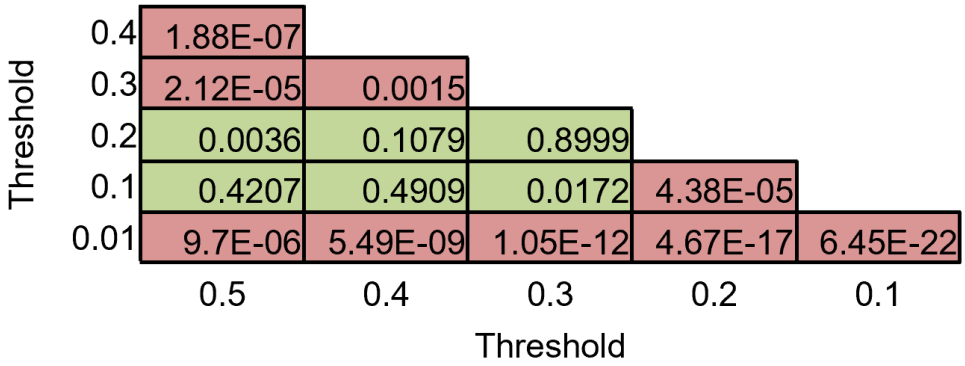
Table 2.2: Absolute percent difference (± standard deviation) of volume, average DSC, and median DSC at six thresholds for the subset analysis. (IQR = interquartile range.)

	Threshold					
	0.5	0.4	0.3	0.2	0.1	0.01
Average absolute % difference of volume	49.48 ± 35.10	42.17 ± 32.89	35.18 ± 30.26	28.70 ± 26.44	23.21 ± 20.32	26.93 ± 18.30
Average DSC	0.61 ± 0.18	0.61 ± 0.17	0.62 ± 0.17	0.62 ± 0.16	0.61 ± 0.15	0.58 ± 0.14
Median DSC (IQR)	0.63 (0.31)	0.64 (0.30)	0.63 (0.27)	0.63 (0.26)	0.62 (0.25)	0.57 (0.23)

Overall, the subset analysis displayed larger volume percent differences and larger standard deviations when compared with values in Table 2.1, although, the average and median DSC values were consistently greater than those acquired when evaluating the entire scan. Lastly, the DSC values achieved from the entire scan and those from the subset were statistically different at the six thresholds evaluated ($p < 0.001$, Figure 2.12). The subset volume was statistically different from the entire scan except for at the 0.2, 0.1, and 0.01 thresholds. Interestingly, the differences between the percent volumes of the entire scan and the subset were smaller at each respective threshold. For instance, the absolute percent difference of volume at the 0.5 threshold was 42.93% and 49.48% for the entire scan and subset, respectively, for an absolute difference of 6.55%. The difference between the entire scan and the subset decreased at the 0.01 threshold, as the entire scan and the subset had percent differences of 26.60% and 26.93%, respectively, resulting in an absolute difference of 0.33%.



(a) P-values for the absolute percent difference of volume.



(b) P-values for the DSC values.

Figure 2.11: Matrix of p-values when comparing the absolute percent difference of volume (a) and DSC (b) across thresholds for the subset analysis. Red indicates a significant difference ($p < 0.0033$ after Bonferroni correction), and green indicates a failure to achieve significance, as determined by the Wilcoxon signed-rank test.

Threshold	DSC: Entire scan versus subset sections	Average absolute % difference of volume: Entire scan versus subset sections
0.5	1.84E-09	1.29E-05
0.4	5.97E-10	1.24E-04
0.3	2.69E-10	0.0022
0.2	2.33E-10	0.0234
0.1	1.52E-09	0.1511
0.01	7.48E-06	0.8261

Figure 2.12: P-values comparing the absolute percent difference of volume and DSC volumes between the entire scan and the subset sections selected. Red indicates a significant difference, and green indicates a failure to achieve significance, as determined by the Wilcoxon signed-rank test. Significance was achieved at $p = 0.0083$, after correcting for six comparisons.

2.3.2 Model calibration using TS

The calculated temperatures for the four validation sets are presented in Table 2.3 below. All temperature values were greater than unity, which demonstrated the model's overconfidence prior to calibration. Figure 2.13 displays an example image during the various parts of the temperature scaling process.

Table 2.3: Summary of the temperatures (which estimate confidence of the model) calculated using the NLL for the four validation sets used in the training process of the model.

Validation set	Temperature
Left tumor plus effusion	3.4
Right tumor plus effusion	2.1
Left tumor only	3.7
Right tumor only	2.3

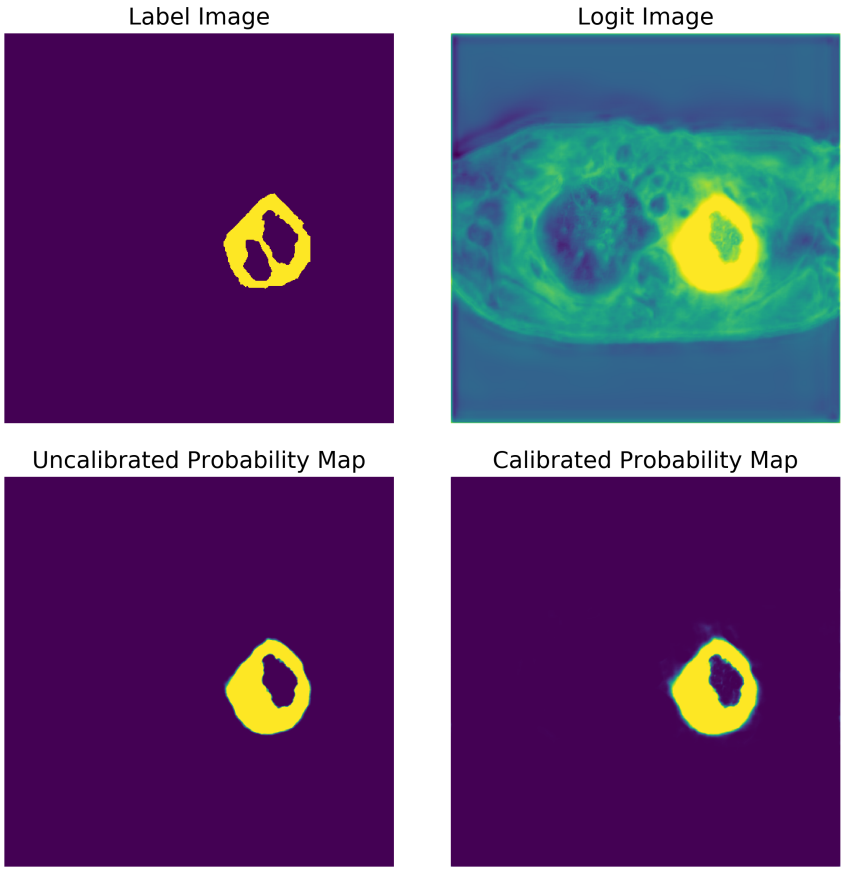


Figure 2.13: Example of a tumor plus effusion case in the left hemithorax at the various stages of post-processing to calculate the temperature. The bottom right image is the calibrated probability vector for the "disease" class, which is the output of the sigmoid activation function of the logit vector $\boldsymbol{z}_i$ scaled to temperature $T = 3.4$.

2.4 Discussion

At lower probability thresholds, more pixels were counted as tumor within a CT section, as demonstrated in Figure 2.2. As a result of a lowered threshold, the computed tumor volume increased (Equation 2.1). Furthermore, except for nine cases, the radiologist's volumes were consistently larger than those of the CNN (Figure 2.6), indicating that the CNN did not fully capture the tumor included in the reference outlines of the radiologist. Lastly, as presented in Table 2.1, the absolute percent difference in volume between the reference contours and the

CNN-derived contours significantly decreased with a lower threshold (Figure 2.4a), indicating an increase of calculated tumor volume using the CNN contours.

While the absolute percent difference decreased at smaller thresholds, the average and median DSC remained within a small range of 0.56–0.59. These values are slightly higher than in the current literature [70]. Therefore, the accuracy of contour overlap did not greatly differ with the change of thresholds. In other words, the newly included pixels at lower thresholds were not at arbitrary locations within the scan but were, on average, at relevant anatomic regions that overlapped the reference contours; this trend demonstrated the robustness of the CNN in identifying tumor, even for lower pixel probabilities. Overall, when inspecting the average of the two figures of merit across all cases, the maximum DSC occurred at a 0.2 threshold, while the lowest absolute percent difference occurred at a 0.06 threshold, as shown in Figure 2.8. A limitation in this analysis was the inherent bias present for the DSC calculation: rather than delineating the tumor on the original CT scans, the radiologist modified the contours generated by the CNN at a threshold of 0.5, which explains the distinct peak at the 0.5 threshold (Figure 2.7).

Six scans from four patients exceeded the 95% limits of agreement (Figure 2.9), and all were underestimates of tumor volume based on the predictions generated by the CNN. The CNN failed to capture disease that surrounded organs such as the spleen, vertebral column, and heart in those scans. For these cases, the radiologist who established the reference standard provided new contours to capture regions of tumor excluded by the CNN at the 0.5 threshold instead of modifying the preexisting CNN output. Severe pleural effusion was erroneously identified as tumor. Overall, these examples underlie major trends of where the model underperformed. Upon closer examination of the images, the CNN had difficulty contouring disease for the aforementioned reasons along with presence of metallic artifacts and disease in the fissure. The underestimation of the CNN-based segmentations is also

captured in Figure 2.5, which shows strong linear correlation with the volume calculated from the reference standard but with a majority of the data above the identity line.

Poor performance was expected for regions that were at the level of the diaphragm or in the lung apices because the model was not trained using such regions. Rather, the model was evaluated using tumor contours applicable to mRECIST measurements. To account for this discrepancy, a subset analysis on sections only between the aortic arch and the pulmonary vein was performed. Contrary to expectation, the percent differences of volume were consistently larger for the subset analysis than for the entire scan across the six thresholds discussed, which warrants further investigation. The difference between the two analyses, however, did decrease at smaller thresholds. Further, the average and median DSC were consistently higher for the subset analysis, which indicate the accuracy of the model when predicting pixels containing tumor. Specifically, the highest median DSC was achieved at a threshold of 0.4 and the highest mean at a threshold of 0.2 for the subset analysis. It is also important to note that the 0.01 threshold yielded the lowest DSC and increased the percent volume difference for the entire scan along with the subset analysis. This indicated that the delineations at low thresholds erroneously identified a substantial number of pixels as tumor. Overall, for both the subset and entire scan, the percent difference of volume decreased with a decrease in threshold, while the average and median DSCs were highest at thresholds 0.2-0.4. Lastly, while volume and DSC distributions seem similar as displayed by the means and medians in Tables 2.1 and 2.2, the values were statistically different as the Wilcoxon signed-rank test calculates the differences of the paired values. Therefore, the paired values changed substantially to yield statistically different results, as shown in Figures 2.4, 2.11, and 2.12.

This work also demonstrated the overconfidence of the initial model, as all calculated temperatures were greater than one. This finding is consistent with the literature, since modern neural networks have been reported to be overconfident in their predictions [6, 7,

10, 13, 14, 15, 17]. With a correctly calibrated model, accurate probability maps can be generated, thus streamlining automation of the mesothelioma segmentation task. Due to the widespread use of neural networks for medical image classification and segmentation tasks, there is a need to ensure that model outputs are properly calibrated so that the resulting probabilities are indicative of the model's true confidence.

There are some limitations that should be addressed in this work. First, as previously mentioned, by displaying the contours that had been generated at the 0.5 threshold to the radiologist, there was an inherent bias: this bias has been shown to impact the modified outlines produced by observers [79]. The 0.5 threshold was chosen a priori as it is often selected because it is an intuitive value to binarize output probability maps. The aim of this study was to determine the impact of the probability threshold on the two figures of merit studied and not to investigate the clinical implementation or the generalizability of a given threshold; therefore, this study was not hindered by having a reference standard with only a single radiologist. Though, the findings of this study could be expanded by using the same two figures of merit to compare the generated outputs of the CNN with a reference standard that is provided independently by multiple radiologists without any computational aid, as that would eliminate this inherent bias. A second limitation was the performance of the segmentation task using a 2D CNN architecture as opposed to more advanced techniques, e.g., a 3D architecture. This will be explored in future work, as the current dataset size may restrict model complexity and result in poor performance when implementing a 3D architecture. Future work will train the CNN using cases for which this current model was deficient, providing it with cases that displayed pleural effusion and disease surrounding the various structures in the thorax.

Overall, the purpose of this work was to study the impact of CNN probability map thresholds on the percent volume differences and DSC values when comparing CNN-generated mesothelioma tumor outlines with radiologist tumor outlines. These results indicate that

the investigation is slightly more nuanced, as lowering the probability threshold (1) predictably increased the resulting tumor volume (which was consistently smaller than that of the radiologist) and lowered the percent difference but (2) only negligibly affected the distribution of DSC values. It is important to make the distinction between CNN volume measurement and DSC. Ensuring that the CNN acquires volume comparable to the reference standard is critical, as volume can be used to capture tumor burden more completely for response assessment. However, similar investigations must also be cognizant of the spatial regions where the CNN identifies tumor; it is not sufficient only to match volumes with the reference standard, but also to match the location of the contours for a more accurate assessment. Thus, investigations concerned with the automated segmentation of PM tumor through similar deep learning approaches need to carefully evaluate the thresholds implemented on the output probability maps as this work has shown the significant differences in tumor volume and spatial overlap with a reference standard as a function of probability map threshold. Future directions of this work will also consider the impact that varied computed tumor volumes have on the tumor response category assigned to patients, enhancing the clinical relevance of this novel work.

2.5 Conclusion

This study explored the impact of changing the threshold applied to the probability maps output by a CNN when segmenting PM tumors on CT scans. After investigating thresholds from 0.001 to 0.9, a clear peak at the 0.5 threshold was found for DSC; however, there was no definitive threshold value to minimize the percent difference of volume between the radiologist and CNN outlines. The percent differences of volume decreased when lowering the probability threshold, while the median DSC values were more robust to threshold changes. The CNN performance was deficient on scans that contained severe pleural effusion and disease that bordered other structures in the thorax. Therefore, a subset analysis was conducted, which

yielded improved results for DSC. Overall, this pilot study highlighted the impact of varying CNN-generated probability map thresholds on mesothelioma tumor outlines, using percent difference of volume and the DSC as the figures of merit.

CHAPTER 3

RADIOMICS FOR DIFFERENTIATION OF SOMATIC BAP1 MUTATION ON CT SCANS OF PATIENTS WITH PLEURAL MESOTHELIOMA

3.1 Introduction

The use of radiomics, specifically texture analysis, has long been implemented in medicine to help clinicians and researchers extract quantitative information from images [19, 21, 22, 23]. Advances in the field have linked imaging features with patients' genetic profiles, i.e., "imaging genomics" [80, 81]. Imaging genomics has been applied to many different diseases and anatomic regions [82]. For example, Velazquez et al. [83] were able to discriminate between cases with and without a somatic mutation in the EGFR gene using radiomic signatures acquired from computed tomography (CT) scans of adenocarcinoma patients. Similarly, Yip et al. [84] performed the same task using positron emission tomography (PET) images of patients presenting with non-small cell lung cancer.

The use of imaging genomics for pleural mesothelioma (PM) is rare in the literature. PM is an aggressive form of cancer present in the pleural lining of the lung, resulting from exposure to asbestos and has a very poor prognosis. The BRCA1-associated protein-1 (*BAP1*) gene encodes for the BAP1 protein, a deubiquitinase that influences cell growth, cell proliferation, and cell death [85, 86, 87]. It is of great interest since it accounts for the most common somatic mutations in PM [87, 88]. *BAP1* mutations can also be inherited, and individuals with germline mutations in this gene have been widely recognized as being predisposed to the disease, although patients with a germline *BAP1* mutation are associated with better prognosis [88, 89] than those without the germline mutation, with a 7-fold increase in long-term survival regardless of sex and age [90]. By identifying suspected germline mutations solely through radiomics, clinicians could be prompted to pursue genetic testing, which is

not currently the standard of care [91], resulting in more streamlined patient prognostication and assessment of family members, who have a 50% chance to inherit the same mutation [89]. To determine the feasibility of future studies determining the germline mutation status from medical images, this novel work[1] first explored the use of radiomics on the CT scans of PM patients with the more-prevalent somatic *BAP1* mutations [90, 93, 94, 95].

3.2 Methods

The overall workflow for this work is presented in Figure 3.1.

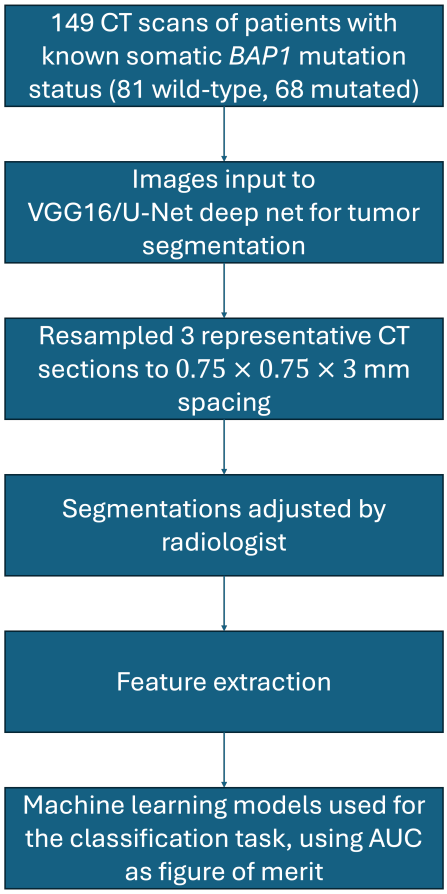


Figure 3.1: Pipeline incorporated in this study, beginning with the patient cohort curated and ending with the machine learning models used for the *BAP1* classification task.

1. This chapter is based on a study reported in [92].

3.2.1 Patient selection and sample collection

This study curated 149 patients diagnosed with PM from the University of Chicago Medicine (UCM) under a Health Insurance Portability and Accountability Act (HIPAA)-compliant, Institutional Review Board (IRB)-approved protocol from April 2016 to June 2022. Informed consent was obtained from all participants. The protocol allowed for the collection and biobanking of peripheral blood, saliva, and tumor samples. Tumor DNA was extracted from fresh frozen, paraffin-embedded tumor tissue blocks. Somatic mutations were identified using the UCM OncoPlus next-generation sequencing panel [96]. Patients with confirmed somatic *BAP1* mutations only (*BAP1*+, N = 68) were included in the study. The remaining 81 patients presented with the wild-type allele (*BAP1*-). Immunohistochemical analysis of the BAP1 protein was conducted in a Clinical Laboratory Improvement Amendments-certified laboratory at UCM using the Santacrus C4 monoclonal antibody. Table 3.1 includes further details about the patients of this study.

Table 3.1: Patient demographics categorizing patient sex and age characteristics.

	Total (n=149)	***BAP1*[+] (n=68)**	***BAP1*[-] (n=81)**
Sex			
Male	95	48	47
Female	54	20	34
Age			
Median	69	69.5	69
Range	21 – 90	51 – 90	21 – 81

3.2.2 Image data curation and segmentation

Axial images from unenhanced chest CT scans of the patients were retrospectively collected (Table 3.2) [97]. Scans were acquired with the assistance of the University of Chicago's Human Imaging Research Office (HIRO) [98, 99], which provided de-identified, compliant images for evaluation. For each patient, the CT section displaying the largest area of tumor

was selected by a radiologist. This section and the immediate superior and inferior sections were used to create a 3D volume for analysis. A Visual Geometry Group 16 (VGG16)/U-Net deep convolutional neural network architecture was utilized to segment the tumor within this volume [43]. The 2D architecture employed a downsampling and upsampling path. The downsampling path utilized a VGG16 model pre-trained on ImageNet with scale-jittering, applying 2×2 max pooling with stride 2. Dropout layers of 0.5 probability were used to prevent model overfitting. The upsampling path employed a 2D operation with nearest-neighbor interpolation on the feature maps. The network generated 512×512-pixel probability maps, which matched the input image size. Rectified linear unit (ReLU) and sigmoid activation functions were applied after the convolutional layers and the final layer, respectively. Lastly, the model was trained with a binary cross-entropy loss function using the Adam optimizer with a learning rate of 10^{-5}. More details regarding the architecture of the model and its training scheme can be found in Gudmundsson et al. [43] and Chapter 2 of this dissertation. For the present study, tumor contours were automatically generated and evaluated with no additional training or validation of the model.

The resulting probability maps output by the network were thresholded at a value of 0.2; this threshold was determined to have maximal overlap with human contours using the Dice similarity coefficient (DSC) from prior work [72, 100]. The radiologist adjusted the resulting segmentations to ensure the segmentations were highly specific to tumor pixels. The finalized contours were defined as the region of interest (ROI) and used for feature extraction.

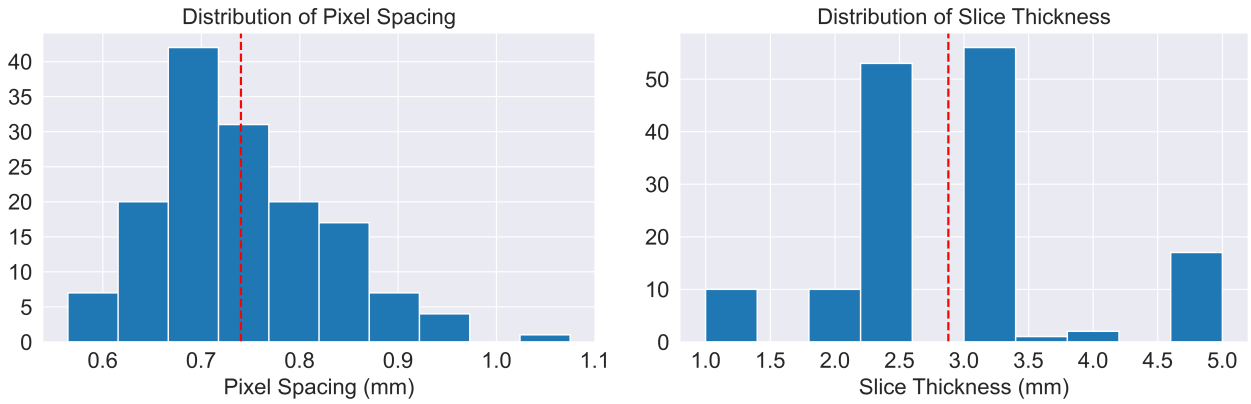
3.2.3 Image resampling and gray-level discretization

To mitigate the impact of different image acquisition parameters, all images were resampled to the mean resolution of all scans, with pixel spacing of 0.75×0.75 mm and a slice thickness of 3 mm (see Table 3.2 and Figure 3.2 for more details). Prior to texture feature extraction, gray-level discretization was applied using a fixed bin number of 32 gray levels, as small

or large gray-level quantizations have been shown to impact texture feature values due to reduction of information that can be extracted from an image [101, 102].

Table 3.2: Image acquisition characteristics for the patient cohort analyzed in this study.

	Total (n=149)	***BAP1*[+] (n=68)**	***BAP1*[-] (n=81)**
Pixel Size [mm]			
Median	0.72	0.71	0.73
Range	0.56 – 1.07	0.57 – 1.07	0.56 – 0.95
Slice Thickness [mm]			
Median	3	2.5	3
Range	1 – 5	1 – 5	1 – 5
kVp [kV]			
Median	120	120	120
Range	80 – 140	100 – 120	80 – 140
Scanner Manufacturer			
GE	73	35	38
Philips	45	21	24
Toshiba	13	5	8
Siemens	18	7	11
Reconstruction Kernel			
GE: Standard	71	34	37
GE: Chest	2	1	1
Philips: B	44	21	23
Philips: C	1	0	1
Toshiba: FC13	6	1	5
Toshiba: FC14	2	1	1
Toshiba: FC18	5	3	2
Siemens: B30f	3	0	3
Siemens: B31f	1	0	1
Siemens: B40f	1	0	1
Siemens: B31s	1	0	1
Siemens: B35s	1	1	0
Siemens: Bf39f	2	1	1
Siemens: Bf37f	1	0	1
Siemens: Br36f	1	0	1
Siemens: Br40d	1	1	0
Siemens: I26f	1	0	1
Siemens: I31f	4	3	1
Siemens: I41f	1	1	0



(a) Distribution of the pixel spacings.

(b) Distribution of the slice thicknesses.

Figure 3.2: Histogram of the (a) pixel spacing and (b) slice thickness of CT sections of the original 149 scans. The red vertical line depicts the mean value in each of the distributions to which resampling was performed.

3.2.4 Feature extraction

Eighteen intensity-based and 123 texture features (111 second-order, six Laws' texture energy, two Fourier, and four fractal dimension) were extracted from the original ROIs. The 123 texture features were also extracted from the ROIs after applying seven different filtering operations on the images: two Laplacian of Gaussian (LoG) filters ($\sigma = 0.75$ mm, 1.5 mm), four multi-channel wavelet decompositions (LH, LL, HL, HH), and a local binary pattern operator (radius = 0.75 mm). With 18 intensity-based features and 123 texture features extracted from the ROIs before the filtering operations and the 123 features extracted from the ROIs after the seven filtering operations, a total of 1,002 features were computed from each ROI (the finalized tumor contours). Intensity-based features were obtained from the 3D volume [103]. All other features were computed by averaging the 2D feature values over the three CT sections. Features were calculated using the Python packages PyRadiomics [104], PyFeats, and Nyxus.

3.2.5 Data imbalance

Due to the imbalance of *BAP1* mutation status among patients, a hybrid approach using the Synthetic Minority Over-sampling Technique (SMOTE) coupled with the removal of Tomek links was employed to over-sample the minority class and under-sample all classes [105], respectively, prior to the feature selection. The SMOTE algorithm generates artificial data in the feature space near existing feature values of cases from the minority class. Tomek links are a pair of nearest neighbors of opposite classes with minimal distance between them compared to other neighboring data. Removal of Tomek links decreases noisy data or eliminates data near the decision boundary. Implementation of SMOTE-Tomek resulted in equal mutation prevalence, per fold, during training.

3.2.6 Machine learning model and feature selection

The performance of 18 separate calibrated machine learning models (Table 3.3) was evaluated using leave-one-out cross-validation (LOOCV), resulting in 149 iterations. Calibration was performed using the "sigmoid" method, which corresponds to fitting a logistic regression model to the scores of a classifier (Platt's scaling). While "isotonic" calibration, which fits a non-parametric isotonic regressor, could be performed, such calibration is recommended only for large datasets as overfitting could result with too few samples (i.e., fewer than 1000 cases) [106, 107].

Feature selection was performed on the training set of each iteration of the LOOCV in the following order (with empirically determined parameters): (1) features with variance less than 0.01 were discarded, (2) features were Z-score normalized, and (3) features with a Pearson correlation coefficient of 0.75 or greater with other features were removed. Lastly, the top four features were selected using the calculated F-value of the analysis of variance (ANOVA) test between the feature and the *BAP1* status. These four features were then extracted from the left-out test case, per iteration, for the classification task.

Table 3.3: Types of models evaluated in the *BAP1* classification task.

Linear
Logistic regression
Ridge
Stochastic gradient descent (SGD)
Passive aggressive
Neighbor
K neighbors
Tree
Decision tree
Extra tree
Support vector machine (SVM)
Linear SVM
Radial basis function SVM
Naive Bayes
Gaussian naive Bayes
Ensemble
AdaBoost
Bagging
Random forest
Extra trees
Gradient boosting
Gaussian process
Gaussian process
Discriminant Analysis (DA)
Linear (LDA)
Quadratic (QDA)

Other training schemes were assessed. In particular, different-sized folds for repeated k-fold cross-validation were implemented as well as changing the number of top features selected. Preliminary work was also performed to study the impact random state seeds had on the classification task.

3.2.7 Evaluation metric and statistical analysis

The receiver operating characteristic area under the curve (ROC AUC) was used as the figure of merit to assess the classification performances of the models to differentiate between *BAP1*+/- patients. The Wilcoxon rank-sum test was used to assess differences in tumor volume and age distributions between patients in the two classes and the DeLong [108] and Wilcoxon signed-rank tests were used to evaluate differences in AUC values between models. To assess the impact of human modifications on segmentation of the PM tumor, DSC values were calculated between the CNN segmentations and radiologist-modified masks to determine the overlap between the two. Further, the classification task was performed employing the same models (Table 3.3) and using feature values extracted from the unmodified CNN probability maps thresholded at 0.2. Using the DeLong test, the AUC values computed from the unmodified segmentations were compared to the AUC values achieved from the modified segmentations. Due to the hypothesis-generating nature of this work, statistical significance was obtained at $p = 0.05$.

3.3 Results

3.3.1 Tumor volume

Figure 3.3 shows the distributions of the tumor volume contoured across the three sections selected per patient in the dataset; the median (range) volume of tumor contoured was 13,109 mm^3 (1,630 – 108,331 mm^3) across all patients, 11,615 mm^3 (1,630 – 108,331 mm^3) for *BAP1*+ patients, and 15,949 mm^3 (1,688 – 92,352 mm^3) for *BAP1*- patients. The difference in median volume between the *BAP1*+/- patients failed to achieve statistical significance ($p = 0.15$), which mitigated the impact of tumor size as a confounding factor for the classification task. *BAP1*+ patients had the larger range of tumor volumes while

BAP1- patients had the larger median. Differences in age between the patient cohorts failed to achieve statistical significance.

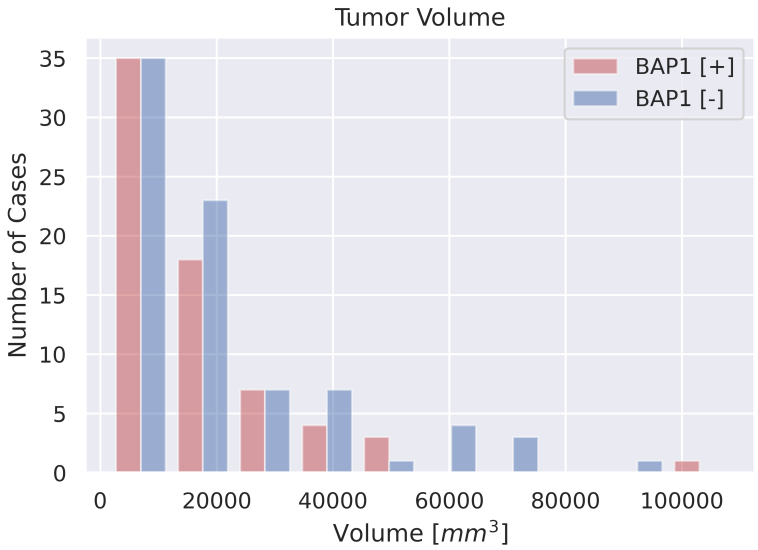
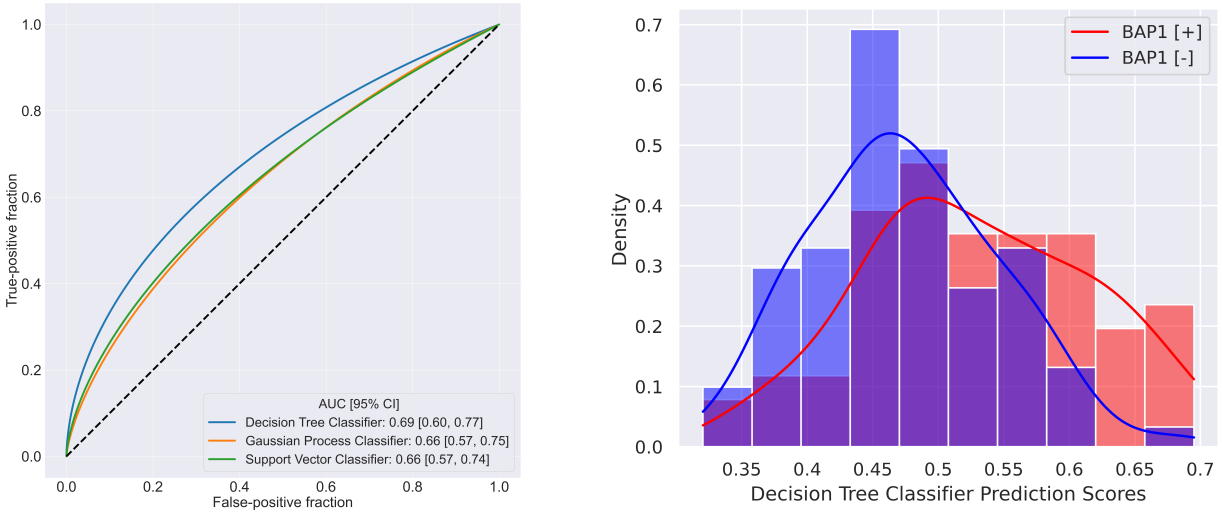


Figure 3.3: Histogram of the tumor volume categorized by *BAP1* mutation status. The difference in tumor volume between wild-type and mutated tumors failed to achieve statistical significance.

3.3.2 Classification performance

In the task of differentiating between *BAP1*+ and *BAP1*- patients, the top three models (sorted by AUC values) were decision tree, Gaussian process, and SVM classifier with a radial basis function kernel (Table 3.3). Figure 3.4a shows the ROC curves obtained from the three models, along with their AUC values and the 95% confidence intervals (CIs). The AUC values and 95% CIs were constructed from 2000 bootstrapped samples of the prediction values during LOOCV. The decision tree classifier yielded an AUC value of 0.69 (95% CI: 0.60, 0.77). Figure 3.4b displays the distribution of scores obtained during the cross-validation for the top-performing model, the decision tree. No scores were less than 0.32 or greater than 0.70 for either class. The DeLong test failed to achieve a statistically significant difference in AUC values among the top three models as shown in Table 3.4.

The four features selected most frequently through the 149 iterations of the cross-validation are presented in Table 3.5. All the features were second-order (e.g., gray level co-occurrence [GLCM] or gray level size zone matrices [GLSZM]) and were extracted from LoG-filtered or wavelet-decomposed images.



(a) ROC curves for the top three models.

(b) Distribution of prediction scores for the decision tree classifier.

Figure 3.4: (a) ROC curves depicting the true-positive and false-positive fractions of the top-three performing classifiers in the task of differentiating somatic *BAP1* mutation status using feature values extracted from segmented regions. ROC curves were fitted using software created by Metz et al. [109]. (b) Distributions of the decision tree classifier prediction scores across all cases. The histograms were normalized to have equal area of one.

Table 3.4: Comparisons of the three best-performing classification models: decision tree, Gaussian process, and support vector. The p-values comparing the differences in AUC values were calculated using the DeLong test, with their corresponding confidence intervals (CIs). Significance levels (α) and widths of the CIs were adjusted for multiple comparisons. None of the comparisons achieved statistical significance after correcting for multiple comparisons using Bonferroni-Holm corrections.

Comparison	p-value for ΔAUC	α	CI of ΔAUC
Decision tree versus Gaussian process	0.4574	0.025	97.5% CI: [-0.060 0.12]
Decision tree versus support vector	0.3208	0.017	98.3% CI: [-0.051 0.12]
Gaussian process versus support vector	0.6478	0.050	95.0% CI: [-0.022 0.036]

Table 3.5: The four texture features most often selected during the 149 LOOCV iterations and the frequency each feature was chosen, i.e., the number of iterations in which a feature was selected.

Transformation	Class	Feature	Selection frequency
LoG ($\sigma = 1.5$ mm)	GLCM	Cluster Prominence	149
LoG ($\sigma = 0.75$ mm)	GLSZM	High Gray Level Zone Emphasis	141
Wavelet (bior1.1–LH)	GLSZM	High Gray Level Non Uniformity Normalized	87
LoG ($\sigma = 0.75$ mm)	GLCM	Correlation	70

3.3.3 Change of k-fold and number of features

Table 3.6 displays the AUC values achieved from the different number of folds used for the repeated k-fold cross-validation and the different number of features selected by the final ANOVA feature selection step: 200 repetitions were performed to ensure robust statistics for the calculation of the 95% CI. As reported in Section 3.3.2, the decision tree classifier resulted in the highest overall AUC value of 0.69 [0.60, 0.77]; however, this AUC value failed to achieve a significant difference ($p = 0.1$) from the AUC value of the SGD classifier (0.63 [0.54, 0.72]) obtained when selecting the top 10 features, as determined using the DeLong test for correlated ROC comparison and setting the alternative hypothesis to "greater."

A selection of four features yielded a different distribution of AUC values than the distribution of AUC values calculated with a selection of 10 features ($p < 0.05$ as determined by the Wilcoxon signed-rank test). There was a significant difference between 10- and 5-fold cross-validation results when selecting four features ($p < 0.05$), however, this trend did not occur for a selection of 10 features as there was a failure to achieve significance ($p = 0.14$). Interestingly, the most-selected feature was the same across all cross-validation approaches: GLCM cluster prominence with an LoG filter applied of size $\sigma = 1.5$ mm (LoG_sigma=2.0). The top-performing models encompassed different types, including ensemble, naive Bayes, discriminant analysis, neighbor, tree, and linear. Therefore, the classification schemes included all but the SVMs and Gaussian processes.

Table 3.6: Model performance using various cross-validation approaches. ROC AUC values in the task of differentiating between *BAP1+* and *BAP1-* patients and 95% CIs for the LOOCV were obtained using 2000 bootstrapped samples. For the 10-fold and 5-fold cross-validation, AUC values were acquired by averaging the AUC values per repeat of the cross-validation approach and 95% CIs were obtained by calculating the 2.5% and 97.5% percentile of the distribution of AUC values.

	Top model	**AUC value [95% CI]**	**Most selected feature**
200-repeat, 10-folds			
Selecting top 4 features	Extra trees classifier	0.58 [0.52, 0.67]	LoG_sigma=2.0 GLCM Cluster Prominence
Selecting top 10 features	Gaussian naive Bayes	0.58 [0.53, 0.62]	LoG_sigma=2.0 GLCM ClusterProminence
200-repeat, 5-folds			
Selecting top 4 features	Quadratic discriminant analysis	0.57 [0.50, 0.64]	LoG_sigma=2.0 GLCM Cluster Prominence
Selecting top 10 features	K neighbors classifier	0.58 [0.51, 0.65]	LoG_sigma=2.0 GLCM Cluster Prominence
LOOCV			
Selecting top 4 features	Decision tree classifier	0.69 [0.60, 0.77]	LoG_sigma=2.0 GLCM Cluster Prominence
Selecting top 10 features	SGD classifier	0.63 [0.54, 0.72]	*LoG_sigma=2.0 GLCM Cluster Prominence

*5 other features were selected during all 149 iterations.

To assess the impact of the random state seed on the performance of a model, AUC values were recorded for 100 seeds of the decision tree classifier, resulting in a median AUC value of 0.66 [0.64, 0.68], with the 95% CI calculated using the percentiles for 2.5% and 97.5% of the distribution of the 100 AUC values calculated; the reported value of 0.69 obtained during LOOCV of the decision tree classifier was outside these boundaries of the CI constructed from the AUC values calculated for the 100 random seeds.

3.3.4 DSC and classification performance of unmodified segmentations

When comparing the CNN segmentations to the radiologist-modified segmentations, an average DSC value of 0.79 with interquartile range of 0.21 was achieved. The same feature extraction and selection was performed on the unmodified segmentations of tumor contours.

The CNN failed to predict tumor for one case, therefore that case was discarded from the analysis. Using LOOCV, the highest AUC value achieved across the 18 models was 0.61. The decision tree classifier, the highest-performing model as aforementioned, yielded an AUC value of 0.45 [0.36, 0.56], which was significantly different than 0.69 ($p < 0.001$) as determined by the DeLong test.

3.4 Discussion

This proof-of-concept work explored the feasibility of differentiating between the mutation status of somatic *BAP1* patients based solely on the 2D radiomics features extracted from patients' CT scans. The approach in this study yielded a higher AUC value than currently reported in the literature (0.65) [63]. To the best of our knowledge, Xie et al. [63] is the only other publication discussing *BAP1* differentiation using image analysis for mesothelioma; however, the work presented here is novel as it is the first to synergistically implement a deep learning model for tumor segmentation and machine learning models for *BAP1* classification.

Prior to the feature extraction, there was careful consideration in the selection of the "standard" reconstruction for all patient scans, attempting to choose this reconstruction across the different scanner manufacturers and kernel nomenclature. In addition, differences in pixel and axial dimension spacing due to variability of image acquisitions from different institutions and different scanners were mitigated by image resampling, as resampling prior to feature extraction has been shown to decrease variability of radiomic features [110]. Similarly, to increase feature stability and reduce noise, gray-level discretization was performed with 32 gray levels [101, 110, 111]. This number of gray levels was chosen based on research extracting features from liver tumor and muscle, but the authors noted that a moderately sized value of gray-level discretization may be applicable to broader radiomic tasks [101]. Future work should consider the optimal discretization employed for this specific work.

After the feature extraction, the classification task was performed through rigorous methodology, employing different machine learning models and cross-validation strategies. It is important to note that different models were evaluated in order to assess the feasibility of this classification task. Further, different models were employed to consider how the different underlying assumptions and parameters of the different models may impact performance. A comparison across the models was also beneficial to ensure that no one model was overfit on the data, resulting in dubiously high AUC values.

LOOCV is known to be a nearly unbiased procedure as the difference in size between the training set in each iteration and the entire dataset is small. There is much discussion about its variability and, more generally, the variance of k-fold cross-validation with different sizes of k. While Efron [112] was one of the first to postulate LOOCV to be unbiased but with high variance, that has since been brought in question [113]. Bengio et al. [114] have shown that no unbiased estimators of the variance of k-fold cross-validation exist. The authors go on to discuss that the variability of LOOCV is impacted by two conditions: (1) if the cross-validation is averaging independent estimates, then LOOCV would return lower variance because of similar reasoning to the low bias as mentioned previously, or (2) if training sets are highly correlated, then LOOCV results in high variance. Overall, LOOCV was chosen a priori because of the small dataset size.

As presented in Figure 3.4a and Table 3.6, the largest AUC value (0.69 [0.60, 0.77]) was achieved using a decision tree classifier when selecting the top four features during LOOCV. The selection of four features was based on preliminary analysis that resulted in moderate performance for classification. However, the AUC value obtained with a selection of four features failed to achieve a significant difference when comparing the AUC value achieved by the SGD classifier and selecting the top 10 features. The 10-fold and 5-fold cross-validation schemes were also implemented to assess the bias and variance of the *BAP1* classification

task. There was comparable performance across the different folds of the various cross-validation methods and different numbers of features selected (Table 3.6).

The most-selected feature obtained using the methodology explained in Section 3.2.6 was the GLCM cluster prominence obtained after application of an LoG filter with radius 1.5 mm. GLCM cluster prominence captures "a measure of the skewness and asymmetry of the GLCM," whereby larger values indicate asymmetry about the mean and smaller values indicate a peak near the mean value and less variation about the mean [104]. The LoG filter first applies a Gaussian kernel to an image, which blurs the image, followed by a convolution with a Laplacian filter (the second derivative of the Gaussian kernel), which enhances the edges in the image. This filter application demonstrated that blurring and enhancing the edges of the ROIs resulted in an appreciable difference between *BAP1+* and *BAP1-* patients that was reflected in the values of the GLCM cluster prominence feature. The other top features (Table 3.5) were either of the GLCM or GLSZM class, both capturing second-order gray-level information about an image. In addition, all four features were selected after application of a filter, three of which were the LoG. It is noteworthy to mention that the only other study that performed radiomics for the *BAP1* classification task reported the relevance of the GLCM cluster prominence feature, as well as the usefulness of other second-order features for classification [63]. Further, the authors found that LoG features were the most stable when extracted from 3D segmentations. Therefore, the findings in this current study support their results.

A comparison between the CNN segmentations and the human-adjusted segmentations was conducted to evaluate the impact human-modified contours had on the classification performance. There was a statistically significant difference between the AUC value obtained from the modified segmentations and the AUC value from the unmodified CNN segmentations. This demonstrated that while this work is the first to combine deep learning for the segmentation task (which substantially reduces the time spent by a radiologist to delineate

the tumor), human input was still required to ensure proper capture of tumor. The increased accuracy of tumor delineations resulted in the moderate performance achieved in classifying *BAP1*+ from *BAP1*- patients.

While this study yielded promising initial results, there are potential future directions in addition to the aforementioned discussion. First, acquiring segmentations on more sections for 3D texture analysis could result in stronger predictive performance by the classifiers as has been reported [63]. Second, stability of the selected features could be assessed through various measures. For example, the concordance correlation coefficient could be used to reduce the number of features based on how well extracted feature values agree before and after image perturbation operations, i.e., rotation or erosion. Initial exploration of stability of features has shown that larger chains of perturbations including rotation and contour randomization produced the most stable and robust feature sets [115, 116]. Third, CT images from the entire history of the patients were visually assessed to identify the scan displaying the largest tumor bulk. Therefore, the selection of scans did control for treatment time point, which could have inherently biased the results, as some of the analyzed scans were acquired either pretreatment or during treatment; the treatment could have potentially affected image features that were extracted if the tissue presented differently. Some scans also were acquired after talc pleurodesis, which could have had an impact on the tumor tissue in a manner similar to that of treatment. Similarly, selecting the section from each scan with the largest visible tumor could have potentially biased the results; as some of these scans had been acquired during the course of treatment, the largest tumor could have been more resilient to the treatment, and the texture features may have captured that resilience as opposed to the mutation status. Future work will expand on this pilot study and select patient scans with stricter criteria, minimizing confounding variables in the curation process. Lastly, while this work demonstrated the ability to accurately detect somatic *BAP1* mutations, the approach will be extended to detect germline *BAP1* mutations in the future.

3.5 Conclusion

The potential of radiomics for identifying *BAP1* mutations from the CT scans of PM patients was demonstrated; 2D features extracted from tumor segmentations yielded an AUC value of 0.69 [0.60, 0.77] when using a decision tree classifier. The novel use of radiomics, machine learning, and deep learning techniques in this work showcased promising results in differentiating between *BAP1*-mutated and wild-type tumors, surpassing previously reported AUC values. While this study showed encouraging outcomes, some future directions are proposed, such as 3D texture analysis, different classification schemes, and assessment of germline mutations.

CHAPTER 4

ASSESSMENT OF A PRE-TRAINED DEEP LEARNING MODEL FOR COVID-19 CLASSIFICATION ON CXRS

4.1 Introduction

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), a ribonucelic acid virus that can impact mammals and birds, is the virus responsible for the ongoing COVID-19 global pandemic. The primary mode of transmission among humans is through exposure to respiratory fluids carrying infectious virus. The virus is highly contagious and can rapidly mutate. Further, infection with the virus may lead to severe or fatal disease. Early detection of the disease can mitigate the symptoms, however, and patient prognosis can improve. Chest radiographs (CXRs) were recommended early in the pandemic for triage, disease monitoring, and assessment of concomitant lung abnormalities (e.g., consolidation, ground-glass opacities, and pulmonary nodules) [53, 54], which resulted in the acquisition of many medical images worldwide. CXRs were also beneficial as they are widely accessible, which makes them an ideal modality for an image-based evaluation of the disease.

With the onset of the COVID-19 pandemic, the artificial intelligence (AI) community quickly joined in the effort to ease the burden on healthcare systems. Before widespread access to reverse transcription polymerase chain reaction (RT-PCR) tests, machine and deep learning (DL) models were developed to provide rapid COVID-19 diagnoses and prognoses based on patients' CXRs and computed tomography (CT) scans [117, 118, 119, 120, 121, 122]. While many DL models reported success in performing these tasks, translating this success to a clinical environment has been difficult due to potential model overfitting or biases present in the datasets [123]. These biases may result in a lack of reproducibility and generalizability of the models developed, a common shortcoming recognized by the AI community [124]. The diversity of images acquired throughout the pandemic, therefore, allows for a comprehensive

evaluation of AI models to assess their generalizability when utilizing the various datasets available.

To determine the robustness of a DL model, an independent dataset can be used along with a performance assessment metric [e.g., area under the receiver operating characteristic curve (ROC AUC)] [125]. If the performance on independent test sets is comparable to the performance on the original test set, then the model is deemed robust. If the model's performance were to decrease, however, then further evaluation is warranted to determine possible deficiencies. While CXRs are not considered a clinical standard for COVID-19 diagnosis, their value lies in their utility for AI assessment. Therefore, the implementation of the DL models in this work is not intended for eventual clinical deployment but rather as a means to thoroughly evaluate the fundamentals of AI as a diagnostic tool.

Overall, the purpose of this work[1] was to validate a deep learning model, using COVID-19 diagnosis from CXRs as the radiologic task, and to compare performance on different datasets while taking into consideration factors such as image-acquisition device (e.g., portable units versus stationary dual-energy subtraction units), patient vaccination status, patient age, and disease severity [127]. These analyses were meant to assess and quantify AI generalizability in the differentiation of COVID-positive and COVID-negative patients based on chest radiography.

4.2 Methods

4.2.1 Datasets

Original dataset

The original dataset consisted of 9,860 patients retrospectively collected from the University of Chicago Medicine as part of an earlier published study [44]. This cohort was split at the

1. This chapter is based on a study reported in [126].

patient level into 64% for training, 16% for validation, and 20% for testing using stratified sampling. COVID-19 prevalence was held constant across the three subsets (15.5%). Only the first CXR exam acquired within two days after a patient's initial RT-PCR test for the SARS-CoV-2 virus was input to the model; RT-PCR tests were used to establish a reliable reference of presence of disease. CXRs had been acquired between January 30, 2020 and February 3, 2021; both standard and soft-tissue images were collected from stationary dual-energy subtraction (DES) radiography units and portable radiography units (these two units define the "CXR exam type"). The portable units generated soft-tissue images using post-processing algorithms. More information regarding this dataset can be found in Hu et al. [44]; the test set from this study will be called the "original test set."

Current test set

Images that comprised the "current test set" were retrospectively collected from 5,893 patients between March 15, 2020 and January 1, 2022 under a Health Insurance Portability and Accountability Act (HIPAA)-compliant, Institutional Review Board-approved protocol. Among these patients, 731 (12.4%) had tested positive and 5,162 (87.6%) had tested negative for the SARS-CoV-2 virus. The current test set served only to assess the performance of the pre-trained model: no additional training or validation was performed. Patient images from the previous study and the current study were acquired from the same institution and were preprocessed in the same manner. Overall, the current test set followed the same curation process as the original dataset to minimize the impact of any confounding variables. A summary of the datasets is presented in Tables 4.1 and 4.2, where Table 4.2 categorizes the three manufacturers that had been used to acquire the standard CXR images for both test sets: Canon Inc., GE Healthcare, and Fujifilm Corporation.

Table 4.1: Summary of datasets used, categorized by various factors, including type of units used to acquire the images.

	Number of patients	Average date of acquisition	COVID prevalence	Number of portable scans (%)	Number of DES scans (%)
Original training set	7,888	08-12-2020	15.4%	6,243 (79.1%)	1,645 (20.9%)
Original test set	1,972	08-13-2020	15.5%	1,595 (80.1%)	377 (19.1%)
Current test set	5,893	03-19-2021	12.4%	4,165 (71%)	1,728 (29%)

Table 4.2: Number of patients categorized by manufacturer and CXR exam type for the original and current test set.

	Original test set		Current test	
	Portable	DES	Portable	DES
Canon Inc.	1,595	12	4,163	43
GE Healthcare	0	359	2	1,596
Fujifilm Corporation	0	6	0	89
Total	1,595	377	4,165	1,728

4.2.2 Image preprocessing

The original Digital Imaging and Communications in Medicine (DICOM) images were gray-scale normalized per image and converted to Portable Network Graphics (PNG) format. Using the converted PNG images, an open-source U-Net-based model was used to segment the lung region from the original dataset and the current test set [128]. The smallest rectangular region that contained the resulting lung mask on a patient's standard CXR image then was cropped; the same mask was applied to a patient's corresponding soft-tissue image. The weights used for the segmentation task were calculated using a pre-pandemic public CXR dataset [129] and further fine-tuned on another dataset of radiographs displaying COVID-19 [44, 130]. Cropping was shown to improve results on the original dataset [44] and was, therefore, performed on the current test set to be (1) consistent and (2) ensure the DL model

would not consider areas outside the lungs (e.g., abdominal region, chest wall, shoulder, and neck region).

The impact of the cropped lung region dimensions on the performance of the classification task was explored on the standard CXRs of the original test set. The U-Net-based model used for lung segmentation initially resized an entire image to 256 $\times$ 256 pixels and cropped the rectangular region that enclosed the predicted lung mask, which will be called the "small lung region." This small lung region was then upsampled to 256 $\times$ 256 pixels by the DenseNet-121 model [131] prior to classification (top panel of Figure 4.1). To study the impact of image resizing, two investigations were performed. First, the U-Net-based model was adjusted to generate the segmented lung region in the same dimensional space as the original image. The resultant "large lung region" image was then input to the DenseNet-121 model for classification (bottom panel of Figure 4.1). The second investigation resized the large lung region image in the bottom panel of Figure 4.1 to the size of the corresponding small lung region image in the top panel of Figure 4.1 before the classification step. All image resizing was performed using the Python PIL package "Image" module, which is also utilized by the U-Net-based model.

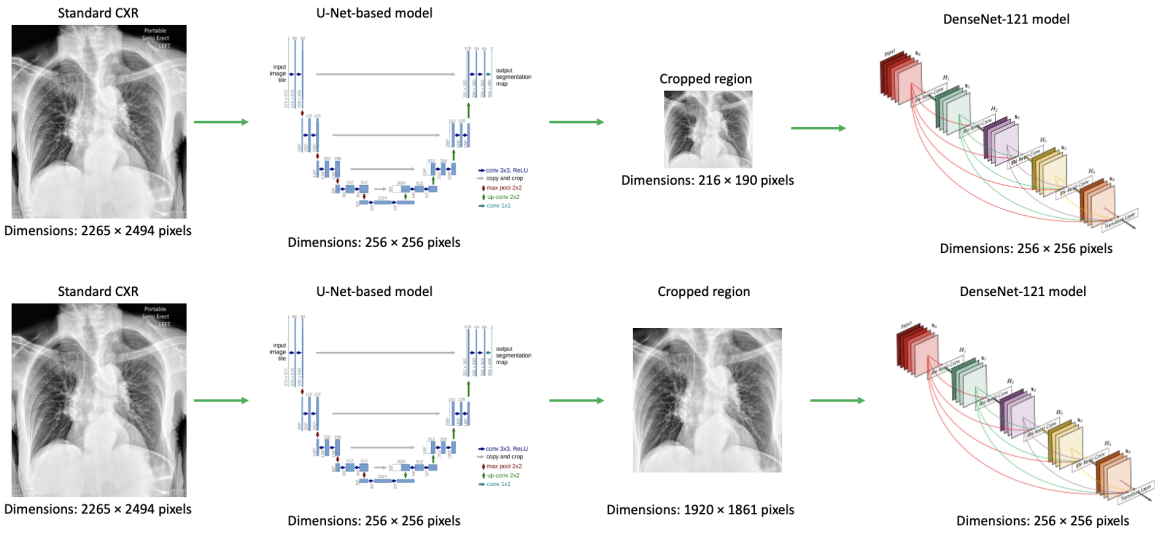


Figure 4.1: The image preprocessing pipeline used throughout this work. Top panel: a patient's standard CXR was resized to 256 × 256 pixels and the lung region was subsequently segmented and cropped, which generated a rectangular region containing only the lung (the small lung region). Bottom panel: after the segmentation task, the cropping was performed in the same dimensional space as the original image. U-Net figure reprinted, with permission, from [74]. DenseNet-121 figure copyright 2017 IEEE [131].

4.2.3 Model implemented

The basis for the present study was a model that was developed by Hu et al. [44] using the original dataset. The model was based on a DenseNet-121 architecture because of its previous success in diagnosing pneumonia and other pathologies on CXRs [132, 133]. Further, it followed a curriculum learning methodology as discussed in Bengio et al. [134]. The curriculum learning became more specialized at each phase, concentrating the model on COVID-19 by the last step. The three phases were: (1) pre-train on ImageNet and fine-tune on the National Institutes of Health (NIH) ChestX-ray14 [135, 136], (2) refine on the Radiological Society of North America (RSNA) Pneumonia Detection Challenge dataset to detect pneumonia [137], and (3) further train on an in-house COVID-19 dataset. Specifically, phase 3 consisted of three classification algorithms developed by image type: standard, soft-

tissue, and a combination of both image types via feature fusion, as presented by Hu et al. [44]. Throughout this study, the algorithms were applied to their corresponding image type (e.g., the algorithm developed with standard CXRs was tested only on standard CXRs). All of the statistical analyses focused on the performance of the current test set when input to the pre-trained model after phase 3, with no additional training or validation performed.

4.2.4 Statistical analysis

Current test set only

Performance of the three classification algorithms was assessed by using images acquired from the different CXR exam types (portable versus DES) on the current test set.

Original test set versus current test set

Comparisons of model performance between the original and current test sets were performed for standard CXRs when considering (1) the entirety of the two test sets (this comparison was repeated for both the soft-tissue CXRs and fusion of the image types), (2) the original test set and only CXRs from the current test set acquired within the date range of the original test set (to control for the different virus strains), and (3) non-immunized patients from both test sets (to control for the impact of disease severity due to vaccines).

The ability of the DL model to classify disease will depend on the severity of the disease as presented on a medical image. Therefore, the COVID severity of the CXRs in the original and current test sets was calculated, as studied by Li et al. [138]. Briefly, the COVID severity model computed a pulmonary x-ray severity (PXS) score, which is defined as the median Euclidean distance between the image of interest and "normal" images (i.e., absence of all pathologies [139]). Specifically, the Euclidean distance, with respect to the imaging features

on which the networks trained, was between the final two layers of twinned DenseNet-121 networks within a Siamese neural network.

The manual modified Radiographic Assessment of Lung Edema scores (mRALE, based on the RALE score created by Warren et al. [140]) were determined by a radiologist with over 20 years of experience on a subset of 50 cases chosen using stratified sampling from the 1,972 original test set standard CXR images (Figure 4.2). The PXS scores were assessed using a Bland-Altman plot [78] to display the agreement between the computed PXS and mRALE scores. Spearman's rank correlation coefficient was calculated to assess the monotonic relationship between the PXS and mRALE scores. Model performance based on PXS score was evaluated after grouping the cases into four equally spaced PXS score bins. Due to the small numbers in the fourth bin for each of the test sets, however, the cases for the third and fourth bins were combined, resulting in three bins for analysis. The "obviousness" of each case was also qualitatively assessed by plotting the DL prediction scores of the cropped standard images from the original and current test sets, with more "obvious" true-negative cases and more "obvious" true-positive cases assigned DL prediction scores closer to zero and DL prediction scores closer to one, respectively.

Uniform manifold approximation projection (UMAP [141]) was used to visualize the penultimate global average pooling layer of the classification algorithm using standard CXR images to qualitatively evaluate the COVID-19 classification task and to visualize the confusion matrix. A quantitative comparison of the two-dimensional UMAPs for the original test set and current test set was performed using one-way multivariate analysis of variance (MANOVA) to test for a significant difference between the two bivariate means of the UMAPs; specifically, the bivariate means were tested for statistical significance using Pillai's trace to calculate the F-statistic, which then resulted in its corresponding p-value. In addition, a comparison of demographics (i.e., age and sex distributions) and International Classification of Diseases 10 (ICD-10) codes between patient cohorts was conducted.

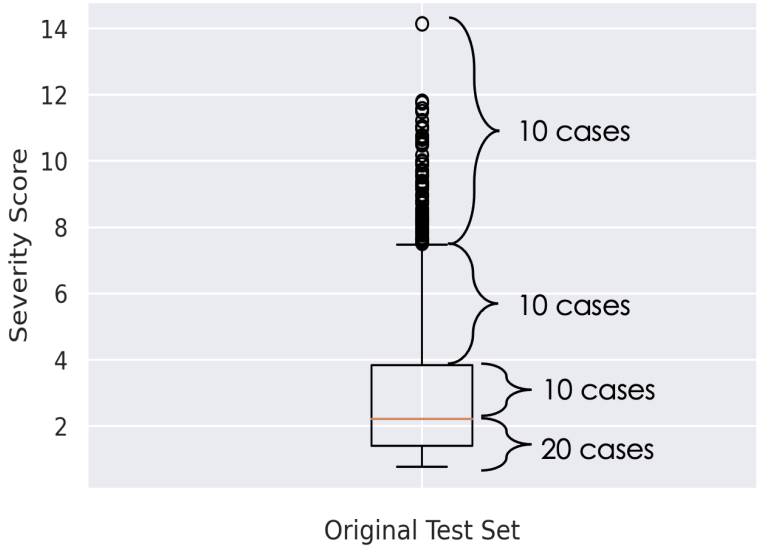


Figure 4.2: Box and whisker plot of the severity score distribution of all 1,972 cases of the original test set. Ten cases were randomly chosen from each part of the box and whisker plot to ensure an equal representation of cases from all possible PXS scores assigned, totaling 50 random cases selected from the original test set to determine the robustness of the PXS scores.

The ICD-10 codes were analyzed to determine whether there were differences in suspected diagnoses between the cohorts that would have resulted in different patient populations. Therefore, varying populations could have explained differences on the CXRs, which would impact model performance. Lastly, an analysis of the performance of the classification algorithm using various standard CXR image dimensions previously described of the image preprocessing steps in Section 4.2.2 was conducted on the original test set.

Overall, these investigations were designed to better understand potential confounding factors that related to image acquisition, different strains of the COVID-19 virus, and patient age. Furthermore, clinical factors, such as vaccination status and severity of COVID were assessed. Performance was evaluated using area under the ROC curve as the figure of merit (2000 bootstrapped samples to construct the 95% confidence intervals), with the DeLong test used to compare the uncorrelated ROC curves [108].

4.3 Results

4.3.1 Current test set only

Classification algorithm

The three classification algorithms corresponded to training the model with the original dataset using (1) cropped standard images, (2) cropped soft-tissue images, (3) and a feature fusion of the two image types. The pre-trained (i.e., no additional training or validation) DL model then was applied to the current test set and achieved the AUC values presented in Figure 4.3.

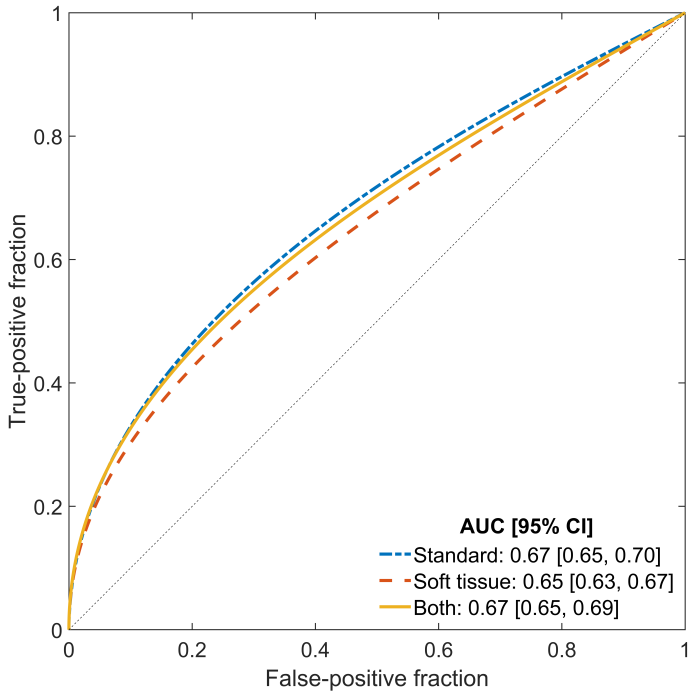


Figure 4.3: ROC curves for the classification of COVID-positive and COVID-negative patients based on CXRs using the current test set. No additional training or validation was performed. ROC curves were fitted using software created by Metz et al. [109].

Table 4.3: Comparisons of model performance for the three classification algorithms: standard images, soft-tissue images, and both types. The p-values comparing the differences in AUC values were calculated using the DeLong test, with their corresponding confidence intervals (CIs) [142]. Significance levels (α) and widths of the CIs were adjusted based on multiple comparisons.

	p-value for ΔAUC	α	CI of ΔAUC
Standard versus soft-tissue	0.0069*	0.017	98.3% CI = [0.0030, 0.050]
Standard versus fusion	0.42	0.050	95% CI = [-0.010, 0.025]
Soft-tissue versus fusion	0.031	0.025	97.5% CI = [-0.039, 0.00075]

*Statistically significant difference after correcting for multiple comparisons (Bonferroni-Holm correction).

The AUC value obtained using the cropped standard CXR images (0.67 [0.65, 0.70]) was significantly higher than that obtained using soft-tissue CXRs (0.65 [0.63, 0.67]). AUC values, however, failed to achieve statistical significance when comparing (1) cropped standard CXRs with both types of images (fusion) and (2) soft-tissue CXRs with both types of images. The results are displayed in Table 4.3.

CXR exam type

When considering the CXR exam type (e.g., portable versus DES unit), the AUC values achieved are shown in Table 4.4. AUC values failed to achieve statistical significance when comparing across CXR exam type for each classification algorithm, i.e., the type of unit did not appear to have an impact on the model's performance in the task of classifying COVID-19.

Table 4.4: Performance of the image type-based classification algorithms for the CXR exam types on the current test set. The 95% CIs are displayed in brackets. Majority of portable images were acquired on Canon Inc. units, and majority of DES images were acquired on GE Healthcare and Fujifilm Coporation units. AUC values failed to achieve a statistically significant difference between exam types for each classification algorithm.

		Portable = 4,165 (70.7%)	DES = 1,728 (29.3%)	Overall = 5,893
COVID-19 prevalence		471 (11.3%)	260 (15.0%)	731 (12.4%)
AUC [95%CI]	Standard	0.68 [0.65, 0.71]	0.69 [0.65, 0.73]	0.67 [0.65, 0.70]
	Soft-tissue	0.66 [0.63, 0.69]	0.64 [0.60, 0.68]	0.65 [0.63, 0.67]
	Fusion	0.68 [0.65, 0.71]	0.68 [0.64, 0.72]	0.67 [0.65, 0.69]

4.3.2 Original test set versus current test set

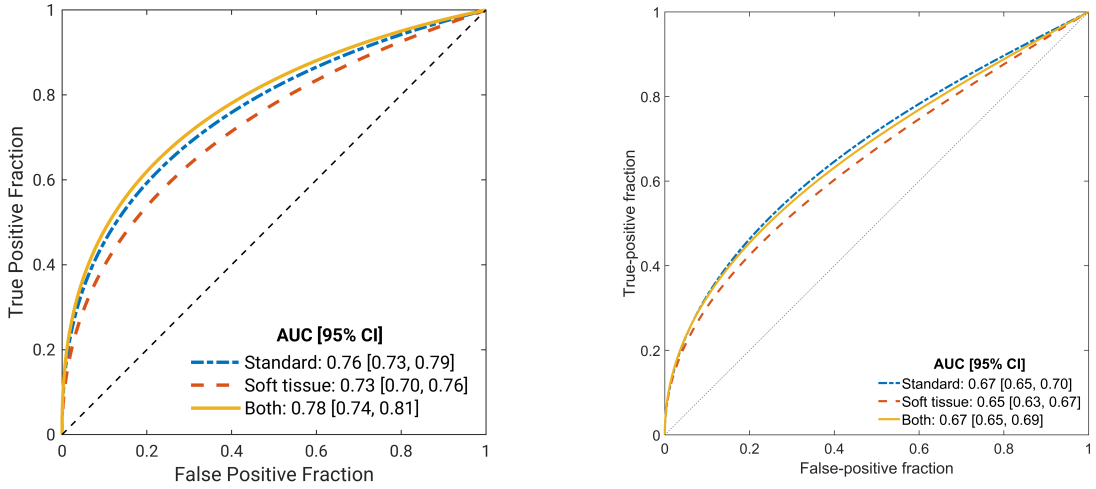
Entirety of both test sets

All AUC values displayed in Figure 4.3 were lower than those obtained in Hu et al. [44] ($p < 0.001$) as calculated using the DeLong test for uncorrelated ROC curves. Figure 4.4 shows the comparison of ROC curves and AUC values between the two test sets.

Date match

There were four main COVID-19 variants of concern (VOC) that underlie the datasets (as defined by the City of Chicago Department of Public Health [143]): the original strain, B.1.1.7 (Alpha) beginning in January 2021 until June 2021, Delta from July 2021 to December 2021, and Omicron from December 2021 onward. Therefore, the original dataset had two VOC: the original variant and Alpha. The current test set had all four VOC. Table 4.5 displays the AUC values calculated when controlling for the variants.

The original and Alpha VOC were controlled for when the current test set was limited to the image acquisition date range of the original test set. An AUC value of 0.66 [0.62,



(a) ROC curves of the original test set for the three classification algorithms.

(b) ROC curves of the current test set for the three classification algorithms.

Figure 4.4: Comparison of ROC curves and AUC values between (a) original and (b) current test sets. AUC values for the three classification algorithms were consistently lower for the current test set compared with the original. Figure 4.4a is the same figure as Figure 6 in ref. [44] (reprinted with permission) and Figure 4.4b is the same as Figure 4.3.

Table 4.5: AUC values calculated for each of the four variants underlying the two test sets.

	Original variant (start to 2020-12-31)	Alpha (2021-01-01 to 2021-06-30)	Delta (2021-07-01 to 2021-11-30)	Omicron (2021-12-01 to present)
Original test set AUC [95% CI]	0.77 [0.73, 0.80]	0.65 [0.52, 0.79]	NA	NA
Current test set AUC [95% CI]	0.67 [0.62, 0.71]	0.68 [0.65, 0.72]	0.71 [0.66, 0.76]	0.63 [0.54, 0.71]
Number of patients in original test set (COVID prevalence)	N = 1,782 (15.9%)	N = 190 (11.1%)	N = 0	N = 0
Number of patients in current test set (COVID prevalence)	N = 1,552 (14.4%)	N = 2,827 (10%)	N = 1,320 (11.6%)	N = 194 (36.6%)
AUC for variant	0.72 [0.70, 0.75]	0.68 [0.65, 0.72]	NA	NA

0.70] (significantly different from the AUC of the original test set, $p < 0.001$) was achieved when considering cropped standard CXRs acquired within the same image acquisition dates as the original test set. This date match corresponds to the overlap of the green histogram bars with the blue as shown in Figure 4.5.

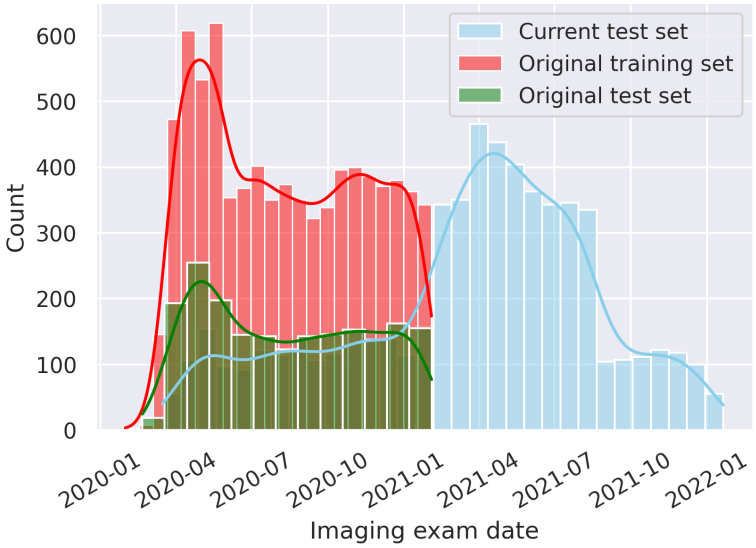


Figure 4.5: Histogram of the imaging exam dates, categorized by the current test set, the original training set, and the original test set. The current test set had a much larger date range, spanning March 15, 2020 to January 1, 2022.

Immunization status

The current test set achieved an AUC value lower than that of the original test set when comparing the cropped standard CXRs of non-immunized patients from the two patient cohorts, as shown in Table 4.6. Further, to determine whether immunization status had an impact on the DL model's performance in classifying COVID-19 status, a comparison between immunized and non-immunized patients in the current test set was performed; differences in AUC values failed to achieve statistical significance ($p = 0.60$).

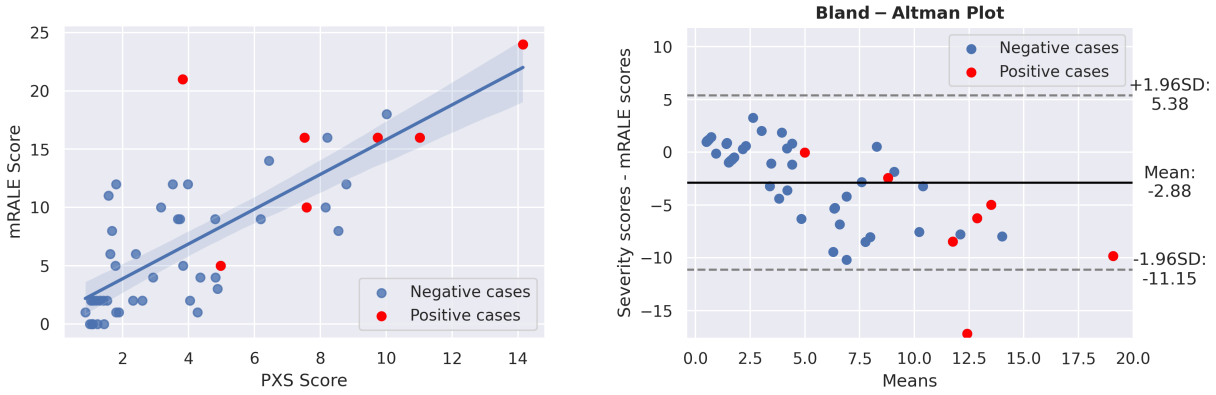
Table 4.6: Summary of statistical analyses performed between the original test set and the current test set. AUC comparisons were conducted using the unpaired DeLong test, and all three comparisons between the test sets were statistically significantly different ($p < 0.001$).

	Original test set	**Current test set**
Total comparison		
Date range	02-20-2020 to 02-03-2021	03-15-2020 to 01-01-2022
Number of patients	1,972	5,893
COVID prevalence	15.5%	12.4%
AUC [95% CI]	0.76 [0.73, 0.79]	0.67 [0.65, 0.70]
Date match		
Date range	02-20-2020 to 02-03-2021	03-15-2020 to 02-02-2021
Number of patients (%)	1,972 (100%)	1,737 (29.5%)
COVID prevalence	15.5%	14.6%
AUC [95% CI]	0.76 [0.73, 0.79]	0.66 [0.62, 0.70]
Nonimmunized patients		
Number of patients (%)	1,966 (99.7%)	4,436 (75.3%)
COVID prevalence	15.5%	14.4%
AUC [95% CI]	0.76 [0.73, 0.79]	0.67 [0.65, 0.70]

COVID severity

COVID severity was assessed for both patient cohorts; differences between the PXS scores failed to achieve statistical significance ($p = 0.17$). PXS scores also failed to achieve a statistically significant difference when matching the image acquisition dates between the test sets ($p = 0.06$). The robustness of the severity score itself was evaluated based on the 50 cases selected in Figure 4.2 using Spearman's rank correlation coefficient ($\rho = 0.74$, $p <0.001$) as well as a Bland-Altman plot to display agreement (Figure 4.6). The Bland-Altman plot showed that the PXS score (calculated using the standard CXRs) was on average lower than the radiologist's assessment (Figure 4.6b).

AUC values resulting from cases in the three PXS score bins for the test sets (details presented in Table 4.7) are displayed in Figure 4.7. The highest AUC value for the original



(a) Scatter plot between mRALE and PXS scores. (b) Bland-Altman plot between mRALE and PXS scores.

Figure 4.6: (a) Scatter plot of a subset of images from the original test set that displays the mRALE score determined by the radiologist and the COVID severity as determined by the DL model described by Li et al. [138]. The regression line is shown in blue where the shaded blue region is the 95% CI. (b) Bland-Altman plot displaying the agreement between the methods of assessing COVID severity. The outlier outside the 95% limits of agreement demonstrated a collapsed left lung with possible effusion [127].

Table 4.7: Definition of PXS score bins for the test sets.

	Original test set	Current test set
PXS score bin edges	0.77 / 4.12 / 7.46	0.76 / 5.12 / 9.48
Cases (N)	1538 / 343 / 91	4914 / 859 / 120
COVID prevalence (%)	11.31% / 26.24% / 45.05%	10.56% / 19.79% / 35.09%

test set was achieved for the third bin, which contained cases with the highest COVID prevalence (45.05%). The highest AUC value for the current test set resulted from the second bin, which contained cases with the second highest COVID prevalence (35.09%): though, the 95% CIs increased with bin edges since the number of cases decreased. The COVID prevalence increased with increasing bin edges, predictably, as higher PXS scores indicate greater radiographic evidence of abnormality. Overall, the original test set yielded AUC values consistently larger than those of the current test set. The distribution of severity scores split according to positive and negative cases, per test set, are shown in Figure 4.8.

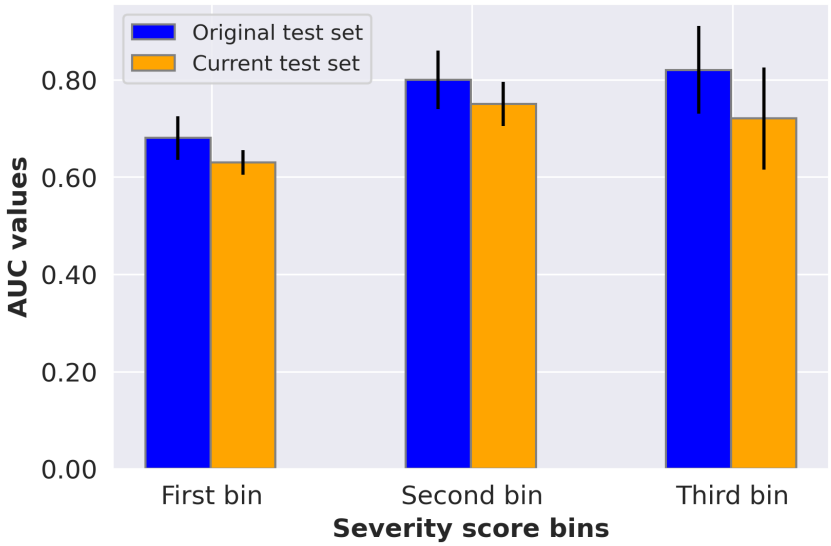
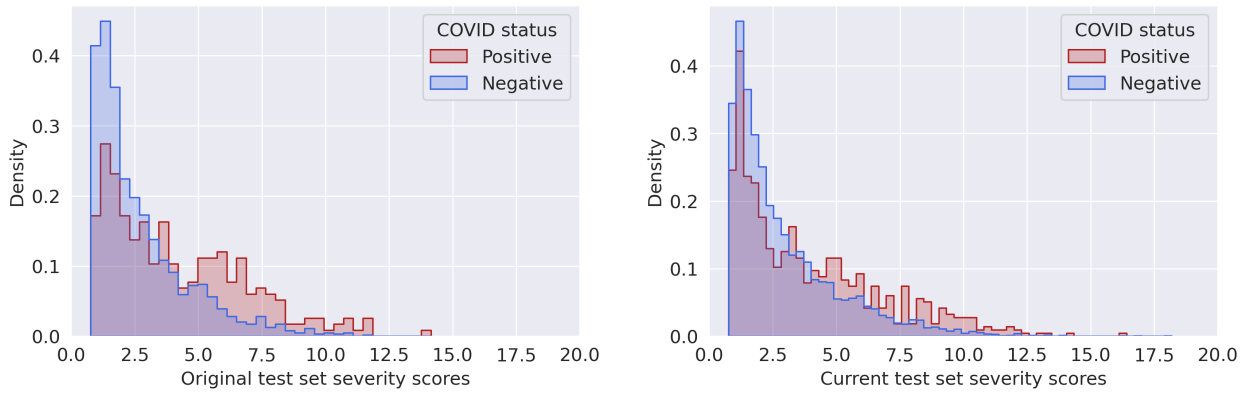


Figure 4.7: Bar plot depicting the resulting AUC values when controlling for PXS scores using the PXS score bin edges defined in Table 4.7. The 95% CIs were calculated by bootstrapping the AUC values 2000 times.



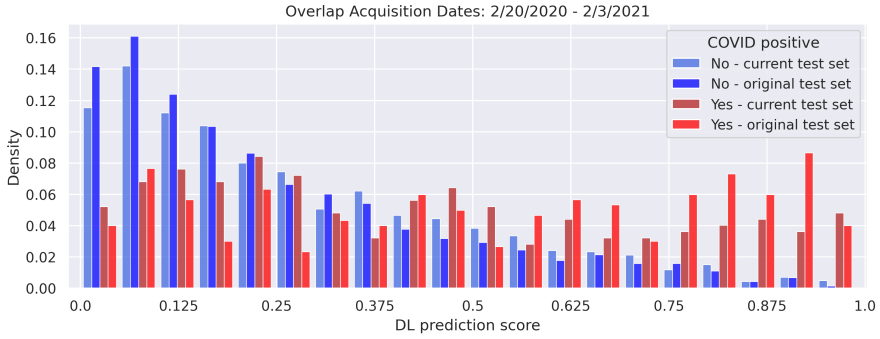
(a) Severity scores of the original test set.

(b) Severity scores of the current test set.

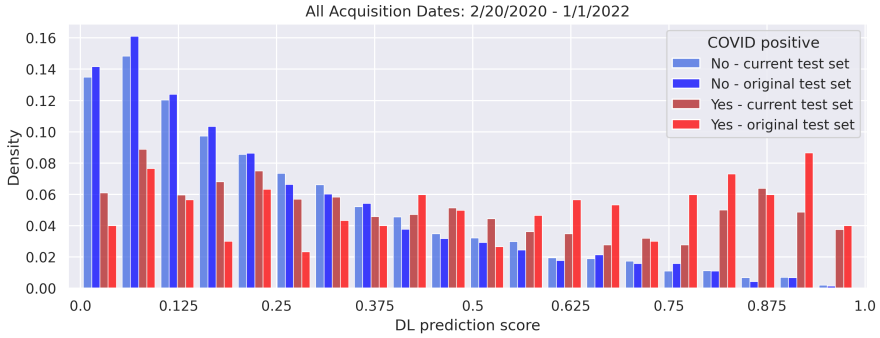
Figure 4.8: Histograms of the severity scores calculated from the original (a) and current (b) test sets. Both distributions demonstrate a strong right skew with higher frequency of positive cases having larger PXS scores.

DL model predictions

Figure 4.9 presents histograms displaying the prediction scores assigned by the model for both the original and current test sets. The distributions display a higher proportion of images from the original test set at low (≤ 0.25) and high (≥ 0.75) prediction scores relative to the images from the current test set, which resulted in the observed higher performance on the original test set. There was a slightly higher count for the current test set at scores in the middle of the plot, i.e., less certain predictions assigned by the DL model.



(a) Prediction scores for the date match.



(b) Predictions scores for the entire image acquisition dates.

Figure 4.9: Histogram of the prediction scores of the DL model for both test sets. The distribution (a) before February 3, 2021, and (b) the entire data range. February 3, 2021 was chosen as the cutoff date as that is the last date which had an overlap of CXR acquisitions between the two patient cohorts (see Figure 4.5). The histograms were normalized to have equal area.

UMAP visualization

UMAP visualizations indicated that the model perceived the two sets of CXRs nearly identically (Figure 4.10). This observation was supported by the MANOVA analysis, which generated an F-statistic of 1.9014 and a p-value of 0.1494, failing to achieve a statistically significant difference between the two bivariate means of the UMAPs generated for the original test set and for the current test set. However, variation existed as the percentages of true positives (TP) and false positives (FP) were different between the original test set (TP = 8%, FP = 11.9%) and the current test set (TP = 4.8%, FP = 12.9%).

(a) 2D UMAP visualization of the global average pooling layer for the original test set.

(b) 2D UMAP visualization of the global average pooling layer for the current test set.

Figure 4.10: UMAP visualization of the confusion matrix for (a) the original test set and (b) the current test set. A similar decision variable was chosen by the deep net for both patient cohorts, classifying positive cases from negative cases (division between blue and orange dots). Overall, the model returned a higher percentage of TPs and lower percentage of FPs for the original test set than for the current test set [127].

Patient demographics

Despite the nearly identical mean and median ages between the two test sets (Table 4.8), the distributions of age (Figure 4.11) yielded statistically significant differences based on the Wilcoxon rank-sum test ($p = 0.006$). There was also a statistically significant difference between ages when matching the image acquisition dates across the two test sets.

Table 4.8: Statistics of the age distributions for the original dataset (which includes training, validation, and test cases) and the current test set.

	Original dataset	Current test set
Mean age (± SD)	54.7 ± 18.9	55.9 ± 19.1
Median age (IQR)	56 (29)	59 (29)

Figure 4.11: Histogram of patient age from the original dataset (which includes training, validation, and test cases) and current test set [127].

There were more men than women in both the original dataset [5088 men and 4772 women (52% male)] and current test set [2979 men and 2914 women (51% male)]. Table 4.9 summarizes AUC values obtained when dividing the test sets with respect to sex. Statistical differences occurred ($p < 0.05$) when comparing the AUC values of each sex of the original test set to the current test set. Differences between sex within each test set failed to achieve a significant difference.

Lastly, ICD codes were compared between the original dataset and current test set, and both sets shared the same top-three ICD-10 codes: (1) encounter for screening for other viral

Table 4.9: Distributions of sex for the original and the current test set with their corresponding AUC values and COVID prevalence.

	AUC value	COVID prevalence
A) Original test set — Male (N = 1051)	0.76 [0.71, 0.80]	14.0%
B) Original test set — Female (N = 921)	0.76 [0.71, 0.80]	17.2%
C) Current test set — Male (N = 2979)	0.65 [0.62, 0.69]	11.1%
D) Current test set — Female (N = 2914)	0.69 [0.67, 0.72]	13.7%

disease (Z11.59), (2) age-related osteoporosis (M81.0), and (3) unspecified osteoarthritis (M19.90). The top three codes are consistent with older patients who were screened for COVID-19. The Z11.59 code was designated for asymptomatic individuals with no known exposure to the virus and either unknown or negative COVID-19 test results [144]. Since the top-three codes were the same, this finding was used as a surrogate to conclude patient diagnoses were unlikely to have accounted for the discrepancy in model performance.

Image dimensions

When using CXRs from the original test set that were resized by the U-Net-based model to the large lung region dimensions (bottom panel in Figure 4.1), the performance of the DL model decreased for all three classification algorithms relative to the published original test set results (Figure 4.4a), obtaining AUC values of 0.74 [0.71, 0.77] for standard CXRs alone, 0.72 [0.69, 0.75] for soft-tissue CXRs alone, and 0.69 [0.66, 0.73] for both types of images. The feature fusion algorithm experienced the largest decrease in AUC value.

To determine the impact of image resizing further, the large lung region images were resized to the small lung region dimensions using the Image module from the PIL Python package. This additional resizing resulted in AUC values closer to those obtained from the original test set analysis: 0.76 [0.73, 0.79] for the standard CXRs alone, 0.72 [0.69, 0.76] for

the soft-tissue CXRs alone, and 0.75 [0.72, 0.79] for both types of images. Once more, the classification algorithm incorporating both types of images experienced the largest decrease in classification performance compared to the 0.78 [0.74, 0.81] AUC value achieved on the feature fusion of the original test set. While the images in this additional analysis were resized to the smaller dimensions using the same Python packages as the U-Net-based model, the pixel values of the cropped images were not identical to those of the original test set. The value of U-Net cropping was demonstrated by using uncropped standard CXRs of the current test set, for which the model achieved an AUC value of 0.58 [0.56, 0.61], substantially lower than the 0.67 reported throughout this work.

4.4 Discussion

The novelty of this study involves its in-depth and exhaustive analysis of various factors that may have contributed to the significant difference in performance by the DL model between the original and current test sets. The results were explored in a variety of ways: performance was assessed across classification algorithm, by CXR exam type, and during different time periods to account for different strains of the COVID virus. Model performance was also evaluated when controlling for equipment manufacturers and various VOC. Vaccination status and disease severity were also considered to determine their impact on the classification task. In addition, patient age and sex were taken into account, along with the model's perception of the radiographs (as captured in the UMAPs) for the two test sets. Influence of the various cropped lung region dimensions was evaluated.

Similar to results obtained from the original test set, the cropped standard CXRs in the current test set performed better when compared with the cropped soft-tissue images. Differences in AUC values between the classification algorithm developed using standard images and the classification algorithm developed using both images (standard and soft-tissue), failed to achieve statistical significance (Table 4.3). AUC values for the three classification

algorithms on the current test set were all significantly lower than those of the original test set.

Dividing the CXRs between portable and DES was performed to investigate whether the type of radiography unit would have an impact on robustness of the DL model as the two types generate images of different quality. A DES unit acquires a soft-tissue image by generating two separate energies of x-rays, creating two images; the resultant soft-tissue image is acquired by subtracting the two images from each other. Portable units generate x-rays only at one energy followed by postprocessing algorithms, which create a synthetic soft-tissue image. Patient geometry is different between the two types of units, as patients are typically oriented in anterior-posterior positioning for portable units and posterior-anterior for DES units. One must also be aware of the motion artifacts that arise from a CXR acquired from a portable unit. Despite these factors, there were no significant differences between CXR exam type AUC values as presented in Table 4.4, unlike the original test set (Table 4 in ref. [44]).

CXR images of the original and current test sets were visually reviewed to ensure no gross differences in imaged patient anatomy, patient positioning, or image artifacts were present between the two test sets; this review did not provide any evidence of systematic differences of this nature. In addition, Gradient-weighted Class Activation Mapping (Grad-CAM) heatmaps were employed to visually assess the predictions of the model for both negative and positive cases from the original and current test sets for high (> 10) and low (< 1) severity as determined by the PXS score. This analysis also did not portray a distinction between the two test sets.

Analysis of patient demographics for the two sets, i.e., age and sex distributions, did not provide any further explanations regarding the discrepancy of performance. While there was a significant difference between age distributions, the findings when matching for sex were consistent with the other investigations: the model better classified COVID-19 status

of patients from the original test set than from the current test set. The ICD-10 codes were also used to characterize potential differences between patient cohorts in terms of suspected diagnoses, but the analysis returned the same top-three codes. Therefore, the CXRs acquired were of similar patient populations.

The original test set had more "obvious" negative and positive cases than the current test set (Figure 4.9), which may have impacted the differences in performance between the two test sets. The "obviousness" of a case was suggested by the increased counts for the low (cases the model perceived as negative) and high (cases the model perceived as positive) prediction scores for the original test set when compared with the current test set (Figure 4.9b). This trend was amplified when limiting the date range prior to February 3, 2021 (Figure 4.9a), i.e., limiting the current test set to the image acquisition date range of the original test set (Section 4.3.2), which likely explains why the date range-matching analysis did not increase the AUC value to one comparable to that of the original test set. Immunization status also did not provide an explanation for the discrepancy in decreased performance.

While the PXS score was designed to evaluate only COVID-positive patients, there was merit in applying this technique to the CXRs of COVID-negative patients because patients obtaining a CXR usually have suspicion of some abnormality in the lung. Therefore, the PXS score, which is defined as the median Euclidean distance between the image of interest and "normal" images, was still a useful metric to incorporate because many of the CXRs of the COVID-negative patients were not "normal."

The data-reduction capability offered by the UMAP was applied to the penultimate global average pooling layer; a two-dimensional embedding was generated that helped visualize how the DL model interpreted the CXRs, thus providing an interpretation of the perception of the radiographs by the model. A nearly identical embedding for images acquired from both cohorts was illustrated. False positive patients appeared to be less concentrated in the embedding for the current test set, however, than for the original test set.

While this work followed the same image preprocessing for the original dataset and current test set, investigation in changes of the cropped lung region dimensions provided insight on how small changes may lead to different results when using DL models. Overall, this work demonstrated the complexity of attaining model robustness and generalizability: an "off-the-shelf" deep net capable of performing classification tasks across different datasets with minimal training remains an elusive task. Therefore, one explanation for the significant difference in performance between the test sets used in this study is a lack of generalizability of the model, which was unable to correctly classify COVID for a new test set (from the same institution) as robustly as it did for the data on which it had been trained originally.

To address this lack of generalizability, future work will investigate (1) the impact of patient demographics and clinical factors on the classifier, (2) whether there were differences in the "obvious" negative or positive patients between the two cohorts, and (3) altering the architecture of the model to make it more robust. First, while patient age and sex was examined for both the original dataset and current test set, matching for age and sex on the training set and test set could provide an explanation for the decrease in performance of the model (e.g., investigate the higher AUC value obtained when considering female patients on the current test set further). COVID severity could also be controlled for between the training and test sets. Second, an analysis of the characteristics of only the positive patients from both cohorts will be performed. For example, if the positive patients in the original test set are older than the positive patients in the current test set, then disease presentation across age may contribute to the performance decrease. Third, a weight regularizer can be applied to the DenseNet layers (specifically, $L2$ regularization) to impose a penalty on the calculated weights, which in turn will prevent model overfitting and possibly make it more generalizable. This will be one approach in potential ablation studies. Repartitioning the original dataset will also be done to determine whether favorable partitions were the reason

for the observed difference in performance. Overall, the regularization could also mitigate the impact of random data partitions, yielding more robust results.

While the the AI community joined in on the efforts early in the pandemic, the community also started recognizing the shortcomings of the methodologies employed, leading to unreliable models [123, 145]. For example, training sets early in the pandemic suffered from small sizes and class imbalances, which made it unlikely that the results of AI models would generalize to broader populations. Goncalves et al. [146] reported that some models originally trained on small datasets from China were intended for use in European populations, resulting in ineffective models due to differences in the three blood biomarkers evaluated in the patient cohorts and the laboratory protocols used. Further, many public datasets of COVID-19 patients comprised images obtained from journal articles without access to the original DICOM images [147], raising concerns about image quality and whether "pictures of pictures" provide the same quality data as original images [145, 148, 149]. What distinguishes this research, however, is the systematic analyses performed to compare datasets that were acquired from the same institution, using the same machinery and imaging protocols; therefore, this novel work provides invaluable insight to how DL models may falter even within the same institution, which in turn can reveal ways to mitigate the lack of model robustness.

4.5 Conclusion

A larger and more current test set of CXRs was used to validate the performance of a pretrained DL model designed to differentiate COVID-positive from COVID-negative patients. AUC values of 0.67 for cropped standard CXRs, 0.65 for cropped soft-tissue CXRs, and 0.67 for both types of cropped images were achieved, which were significantly lower than performances of 0.76, 0.73, and 0.78, respectively, on the original test set. Several factors were considered to determine their impact on the observed differences in model performance

on the test sets, including time period of image acquisition, immunization status, age and sex distributions, and disease severity. The underperformance of the model on the current test set may be explained by a lack of model generalizability. Overall, this research highlighted the importance of not only developing DL models but also rigorously testing their performance across various scenarios to ensure robustness and generalizability.

CHAPTER 5

CONCLUSIONS AND FUTURE DIRECTIONS

This dissertation extensively researched the implementation of deep learning (DL) methods for the segmentation of pleural mesothelioma (PM) on computed tomography (CT) scans and classification of COVID-19 on chest radiographs (CXRs). Further, machine learning (ML) techniques were combined with DL to perform imaging genomics through the use of radiomics and texture feature analysis.

Chapter 2 employed a DL algorithm, namely a Visual Geometry Group 16 (VGG16)/U-Net model [74, 75], to automatically segment PM tumor as presented on CT scans acquired as an independent and external dataset. The ability of the model to segment the tumor was evaluated using two figures of merit: the percent difference of volume and the Dice similarity coefficient (DSC). The two metrics were calculated between the predicted segmentations and a reference standard. Tumor volume was quantified as it provides a more accurate assessment of tumor extent and response to therapy [64, 65]. This work in particular aimed to quantify the impact various probability thresholds of the generated segmentations have on the two figures of merit. No single threshold for the CNN probability maps was optimal for both tumor volume and DSC. This work, however, underscored the need to assess tumor volume and spatial overlap when evaluating CNN performance. While automated segmentations may yield comparable tumor volumes to that of a reference standard, the spatial region delineated by the CNN at a specific threshold is equally important.

Chapter 3 presented the first investigation of the use of both DL and ML algorithms for the automatic segmentation of PM and classification of somatic *BAP1* mutation on CT scans, respectively. The same DL model discussed in Chapter 2 was used to segment PM tumor on three representative CT sections per patient. The generated segmentations were adjusted to ensure high specificity of pixels depicting tumor. Texture features were then extracted from the resultant segmentations and used for the classification task. Using ML models, a

decision tree classifier was able to yield moderate performance in the task of differentiating between *BAP1*+/- patients. The findings of this study can be leveraged for future germline *BAP1* mutation research; germline *BAP1* mutations are more clinically relevant as family members have a 50% chance to inherit the same mutation [89], and germline testing is not commonly performed [91]. Therefore, identification of genetic information through image analysis could lead to improved patient prognostication and family member assessment.

Chapter 4 continued the DL investigations by assessing the performance of a pre-trained DenseNet-121 model [131] in the task of classifying patients as COVID+/- based on CXRs while considering various image acquisition parameters, clinical factors, and patient demographics. Performance of the model trained using standard and soft-tissue CXRs of an original dataset was compared to the performance of the same model on a larger more-current test set. The current test set contained a larger span of dates, incorporated different variants of the virus, and included different immunization statuses. Model performance on the current test set was significantly lower than the performance of the model on the original test set. Investigations that matched the acquisition dates between the original and current test sets (i.e., controlling for virus variants), immunization status, disease severity, and age and sex distributions did not fully explain the discrepancy in performance. Therefore, the lower performance on the current test set may have occurred due to model overfitting and a lack of generalizability.

Future work extending the research of this dissertation can address some of the limitations posed. The following paragraphs will discuss the limitations and potential future directions for the separate chapters presented.

While the VGG16/U-Net model used in Chapter 2 achieved initial strong performance [42, 43], there are more recent and advanced models that have been developed for medical image segmentation tasks [150, 151, 152]. One potential future direction to help improve the mesothelioma segmentation task could be to implement a region-based fully convolutional

network (R-FCN), which performs instance segmentation instead of semantic segmentation (Section 1.3.1) [150, 152]. This approach can perform multi-region segmentation by localizing the regions of interest and subsequently performing binary classification for every region separately. For the mesothelioma segmentation task, the regions of interest may be mesothelioma tumor and adjacent atelectatic lung tissue. Identifying the two regions as separate tasks could improve performance as the regions have similar visual characteristics on CT scans. R-FCNs can be implemented using cascaded FCNs, whereby one FCN is "stacked" on another, with the former used to locate the region of interest and the second performs the classification task [152]. Separately, U-Nets coupled with generative adversarial networks (U-Net-GANs) can be used for probability map generation and discriminators, respectively.

Other possible future directions for the mesothelioma segmentation task would involve implementation of the V-Net architecture, a 3D version of the U-Net architecture for medical image segmentation [150, 151, 152]. The V-Net uses a loss function based on the DSC (instead of binary cross-entropy loss), which is beneficial when there is an imbalance of pixels labeled as tumor and those labeled as background. Unlike U-Nets, V-Nets use residual blocks as short skip connections between shallower and deeper convolutional layers, which improves the convergence when compared with U-Nets [152]. However, the disadvantage of using 3D architectures is that the dataset size used for training is substantially reduced because model training is now performed at the patient level rather than at the image level. One approach in the literature to mitigate this limitation is to implement a 2.5D FCN segmentation [153]. In that work, the authors employed image patches consisting of several consecutive axial slices, which were used as inputs to the FCN and used a "majority voting scheme" for segmentation. The authors also note the importance of transfer learning, as their FCN was able to achieve superior performance. Lastly, while mainly academic, model calibration should be considered, in particular, when evaluating the generated segmentation maps and the probabilities output by the DL algorithms.

Chapter 3 demonstrated the ability of radiomics, specifically in the context of imaging genomics, to distinguish between somatic *BAP1*-positive and *BAP1*-negative patients based on the extracted texture features of the tumor regions of interest. Further, no prior study has combined a DL model for segmentation and ML algorithms for the classification in this particular task; however, there are limitations that hinder reaching a strong conclusion about the feasibility of this task. For instance, the small dataset size, especially compared with the number of features extracted, may have rendered the problem an underdetermined system. Further, the stability and robustness of the features were reduced as features were extracted from only three CT sections per patient, and the CT scans had been acquired across a wide range of clinics, scanner manufacturers, and at different time points during the course of treatment for the patients. Future investigation should benefit greatly from a larger patient cohort to reach stronger conclusions. The methodology established in this work, however, should translate over to the increased dataset size.

Chapter 4 provided a comprehensive evaluation of a DL model trained to classify COVID-positive and COVID-negative patients based on their CXRs. While the work was extensive, no real-world characteristics, such as age, sex, image acquisition dates, and COVID severity were able to explain the discrepancy in model performance between the original dataset and the current test set. Therefore, future work can pivot toward analyses of model retraining and data partitioning instead of patient and clinical information. By examining the impact of different training schemes and data partitions, these investigations may provide a more complete assessment of this deep CNN, as well as DL algorithms in general. Having access to two, in-house, independent datasets provides a more-controlled opportunity to examine the data and the model. This is particularly beneficial as there are fewer confounding factors to account for when investigating the discrepancy in model performance, in contrast to using publicly available image repositories, for example. Overall, the concept of model

generalizability is further addressed in the Appendix, as it compares performance of the model on a sample (original dataset) and target (current test set) population.

This dissertation explored the application of DL and ML methods for the segmentation of PM on CT scans, classification of *BAP1* mutation using imaging genomics, and the classification of COVID-19 on CXRs. For the mesothelioma segmentation task, a VGG16/U-Net model was used, and the impact of various probability thresholds on tumor volume and DSC was evaluated. The work emphasized the importance of assessing both tumor volumes and spatial overlap when evaluating DL model performance compared with a reference standard. Using the same model, PM tumor was automatically segmented from another patient dataset, and texture features were extracted and used to successfully classify somatic *BAP1* mutation on the basis of CT scans. These initial findings suggest the potential application of texture feature analysis to patients with germline *BAP1* mutation, leading to improved patient prognostication if successful. Lastly, for the COVID-19 work, a pretrained DenseNet-121 model was used for CXR classification, which resulted in a significant decrease in performance on a more current test set, possibly due to model overfitting and lack of generalizability. The findings of the COVID-19 study urge further exploration into model retraining and data partitioning for the enhancement of model generalizability. Overall, this dissertation contributes valuable insights into medical image analysis using DL and ML, paving the way for potential advancements in mesothelioma segmentation and *BAP1* and COVID-19 classification through image analysis.

APPENDIX A

IMPACT OF MODEL RETRAINING ON A DEEP LEARNING MODEL IN THE TASK OF COVID-19 CLASSIFICATION ON CXRS: A PILOT STUDY

A.1 Introduction

In the early stages of the coronavirus disease 2019 (COVID-19) pandemic, deep learning (DL) algorithms emerged as potential tools for rapid diagnosis of the virus based on the chest radiographs (CXRs) of patients. As the deployment of these algorithms progressed, however, it became evident that their performance was not always consistent, and challenges arose in ensuring their reliability in clinical settings. For example, some models were trained on CXRs of pediatric patients but were then applied to an adult population, which resulted in models predicting whether the patient was a child, not COVID-19 status [123]. Similarly, a model trained on images of patients lying down and standing up was able to identify the status of patient positions, instead of disease status, with the intuitive notion that patients lying down were more likely to be ill [154]. Further, most models struggled with robustness and generalizability, as there was poor truth labeling that sometimes relied on subjective assessments by physicians rather than more objective metrics such as reverse transcription polymerase chain reaction (RT-PCR) tests [123]. Data collection was also a hindrance, as some available public datasets amalgamated data from various sources that may have included duplicate images, resulting in some CXRs being used in both the training and test sets, which yielded overly optimistic results [123, 154]. In all, a majority of models assessed early in the pandemic were not ready for clinical deployment, as there were inherent biases present [123, 155].

There were early efforts to combat data biases and lack of model generalizability. For example, to structure data curation, the Medical Imaging and Data Resource Center (MIDRC)

was created with an aim "to foster machine learning innovation through data sharing for rapid and flexible collection, analysis, and dissemination of imaging and associated clinical data by providing researchers with unparalleled resources in the fight against COVID-19" [156]. Further, MIDRC conducted a grand challenge to assess performance and generalizability of DL models in the task of distinguishing between COVID-19 positive/negative CXRs [157]. Thus, the present study, along with previous work presented in Chapter 4 [126], was motivated by examining model generalizability. Specifically, a previously published DenseNet-121 DL model obtained an area under the receiver operating characteristic curve (ROC AUC) value of 0.76 in the task of COVID-19 classification [44]. When employing the same pre-trained, original model on an independent test set from the same institution, a significantly lower AUC value of 0.67 was achieved [126]. Therefore, the motivation of this work was to investigate the discrepancy in performance, and lack of generalizability, of the original model applied to the two test sets acquired from the same institution.

As DL algorithms become more widespread in healthcare tasks, it is imperative that artificial intelligence (AI) scientists can understand and interpret the outputs of these models to explain and mitigate potential inconsistencies in model performance. Overall, this current study aimed to provide an interpretation for the outputs of the DL model in question, addressing the discrepancies between these two datasets by examining data partitioning, model architecture, and training, in an effort to understand the lack of model generalizability. This research aimed to contribute to the ongoing efforts in improving machine learning performance for COVID-19 diagnosis and other radiologic tasks, which could benefit from AI deployment in a clinical setting.

A.2 Methods

A.2.1 Datasets

Set A

Set A included 9,860 patients retrospectively collected from the University of Chicago Medicine under a Health Insurance Portability and Accountability Act (HIPAA)-compliant, Institutional Review Board (IRB)-approved protocol. The dataset was initially partitioned into 64% for training, 16% for validation, and 20% for testing using stratified sampling to maintain a consistent COVID-19 prevalence of 15.5% across the subsets. This training and validation set will be termed Set A_{tr}, and the test set will be termed Set A_{te}. Only the first CXR image acquired within two days of a patient's initial RT-PCR test for the SARS-CoV-2 virus was used. CXRs were acquired between January 30, 2020 and February 3, 2021 using standard images from stationary dual-energy subtraction radiography units and portable radiography units. For further details on this dataset, refer to Hu et al. [44].

Set B

CXR exams collected from 5,893 patients constituted Set B and had been acquired between March 15, 2020 and January 1, 2022, under the same HIPAA-compliant, IRB-approved protocol. Within this cohort, 731 patients (12.4%) had tested positive, while 5,162 patients (87.6%) had tested negative for the SARS-CoV-2 virus, as determined by RT-PCR tests. Patient images from both Set A (the initial set used to develop and evaluate the published model) and Set B (the newer set used to evaluate the published model) were obtained from the same institution and underwent identical image preprocessing. The curation process for Set B paralleled that of Set A to mitigate the impact of potential confounding variables. For further details on this dataset, refer to Chapter 4 and Shenouda et al. [126]. Table A.1 provides an overview of the two datasets.

Table A.1: Number of patients and COVID prevalence for Set A and Set B.

	Number of patients	COVID prevalence
Set A_{tr}	7,888	15.4%
Set A_{te}	1,972	15.5%
Set B	5,893	12.4%

A.2.2 Image Preprocessing

Digital Imaging and Communications in Medicine (DICOM) images of the CXR exams were gray-scale normalized and converted to Portable Network Graphics (PNG) format on a per-image basis. Subsequently, an open-source U-Net-based model [128] was used to segment the smallest rectangular region containing the lungs on the PNG images from both Set A and Set B. The segmentation model weights were computed using a pre-pandemic public CXR dataset [129] and fine-tuned on another dataset featuring COVID-19 radiographs [44, 130]. Cropping was performed as it was shown to be effective for the original model [44] and to maintain a consistent methodology.

A.2.3 Model Training Scheme

The current study is based on a model described by Hu et al. [44], which used a single, distinct partition of Set A: Set A_{tr} for the training and validation sets and Set A_{te} for the test sets. The model employed a DenseNet-121 architecture [131], chosen for its previous success in diagnosing pneumonia and other pathologies on CXRs [132, 133]. Additionally, it adopted a curriculum (transfer) learning approach [134], increasing the focus of the classification task towards COVID-19 in the final phase. The curriculum comprised three phases: (1) fine-tune the model pre-trained on ImageNet on the National Institutes of Health (NIH) ChestX-ray14 dataset [135, 136], (2) refine on images from a pneumonia detection challenge [137], and (3) further fine-tune using the initial partition of Set A split into Set A_{tr} and Set A_{te}[44].

A.2.4 Analyses and Comparisons

Using Set A_{tr} to train and validate, the original model yielded an AUC value of 0.76 [0.73, 0.79] (2000 bootstrapped samples to construct the 95% confidence intervals) on Set A_{te} in the task of distinguishing COVID+/- patients from their cropped standard CXRs. Using the same pre-trained model, Set B yielded an AUC value of 0.67 [0.65, 0.70] (also calculated from 2000 bootstrapped samples), which was significantly lower than the results of Set A ($p < 0.001$) as determined by the DeLong test comparing the uncorrelated ROC curves (Figure A.1) [108]. To investigate the decrease in model performance from Set A to Set B, the present study investigated different model retraining strategies, an ablation technique, data partitioning, and model deployment on a grand challenge to assess model performance.

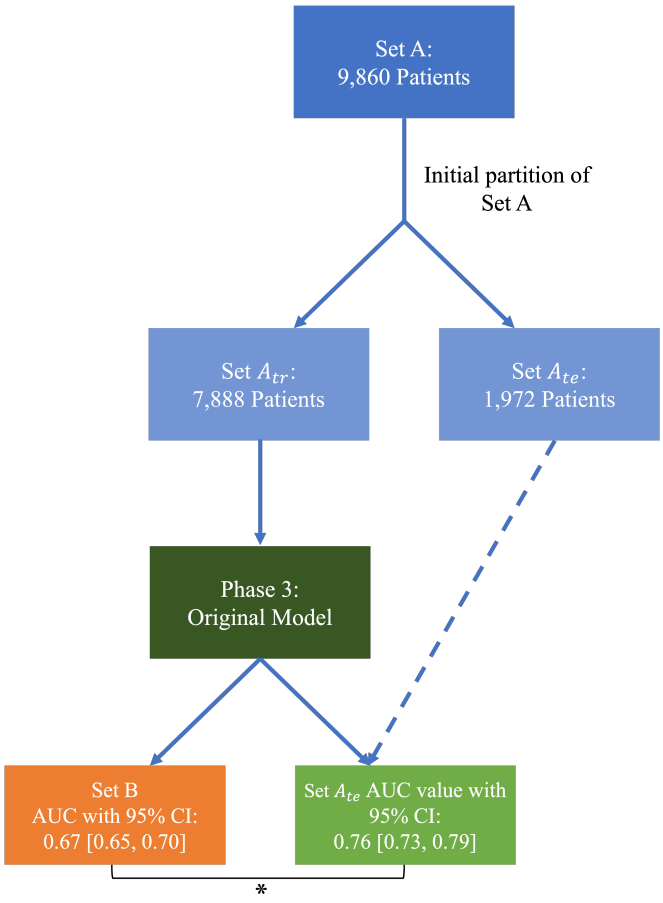


Figure A.1: The comparisons performed between the initial partition of Set A and Set B. The AUC value refers to the test set of Set A, Set A_{te}. The asterisk denotes the statistically significant difference between Set A_{te} and Set B. The green and orange boxes indicate results on Set A_{te} and Set B, respectively.

The first investigation, Experiment I, used Set B to retrain the model to calculate new phase 3 weights by employing the same split ratios as Set A: Set B was split into 64% training (3,771 patients), 16% validation (943 patients), and 20% testing (1,179 patients), using the cropped standard CXRs and maintaining the COVID prevalence at 12.4% across partitions. The combination of the Set B training and validation partitions will be termed Set $B_{tr,I}$ and the Set B test set will be termed Set $B_{te,I}$.

The second investigation, Experiment II, independently fine-tuned the model after the original phase 3 was conducted. Specifically, Set B was used to fine-tune the model after

the original phase 3 weights by splitting the set into 40% training (2,356), 10% validation (590 patients), and 50% testing (2,947 patients). The combination of the Set B training and validation partitions will be termed Set $B_{tr,II}$ and the Set B test set will be termed Set $B_{te,II}$.

An ablation study, Experiment III, was also performed by altering the architecture of the original model for phase 3. Specifically, an $L2$ regularizer (with an $L2$ regularization penalty of 0.0005) [158] was added to help mitigate overfitting, constraining the complexity of the model by minimizing the values the learned weights can take during phase 3. This was performed using the initial partition of Set A (Set A_{tr} and Set A_{te}).

For Experiment IV, the phase 3 weights were recalculated for each of 200 repartitions of Set A_{tr}, and each of the resulting 200 models was evaluated on Set A_{te} and Set B to quantify impact of data partitioning on performance. Specifically, the training and validation sets that comprise Set A_{tr} were separately resampled with replacement 200 times. These 200 partitions will be termed Set $A_{tr,IV}$ (200 instantiations of Set $A_{tr,IV}$ were generated).

Lastly, the original model was also evaluated during the validation phase of the Medical Imaging and Data Resource Center (MIDRC) COVIDx Grand Challenge, which was conducted in November 2022, to determine the performance of the model on a dataset outside the institution on which the model was originally trained and tested. No additional training or validation was performed for the evaluation on the images from the grand challenge. All data partitions employed stratified sampling to maintain a consistent COVID-19 prevalence.

Comparisons of model performance between Sets A and B were performed for standard CXRs when considering the four experiments: (1) recalculating the phase 3 weights using Set $B_{tr,I}$, (2) fine-tuning the phase 3 weights using Set $B_{tr,II}$, (3) implementing the $L2$ regularizer on the original model and retraining the phase 3 weights, and (4) repartitioning Set A_{tr} 200 times and recalculating the phase 3 weights, thereby evaluating whether the

initial Set A results on Set A_{te} were due to an initial chance favorable partitioning. Table A.2 and Figure A.2 summarize the methods and comparisons performed.

Table A.2: Summary of the datasets used and comparisons performed. Of note, cases from the MIDRC Grand Challenge were assessed using the original model, which was pre-trained on the initial partition of Set A, Set A_{tr}.

Experiment	Strategy or application	Training set	Comparison
I	Recalculating phase 3 weights	Set $B_{tr,I}$ (N = 4,714)	Set A_{te} (N = 1,972) and Set $B_{te,I}$ (N = 1,179)
II	Fine-tuning phase 3 weights	Set $B_{tr,II}$ (N = 2,946)	Set A_{te} (N = 1,972) and Set $B_{te,II}$ (N = 2,947)
III	$L2$ regularization applied during phase 3	Set A_{tr} (N = 7,888)	Set A_{te} (N = 1,972) and Set B (N = 5,893)
IV	200 repartitions and recalculating phase 3 weights	Set $A_{tr,IV}$ (N = 7,888)	Set A_{te} (N = 1,972) and Set B (N = 5,893)

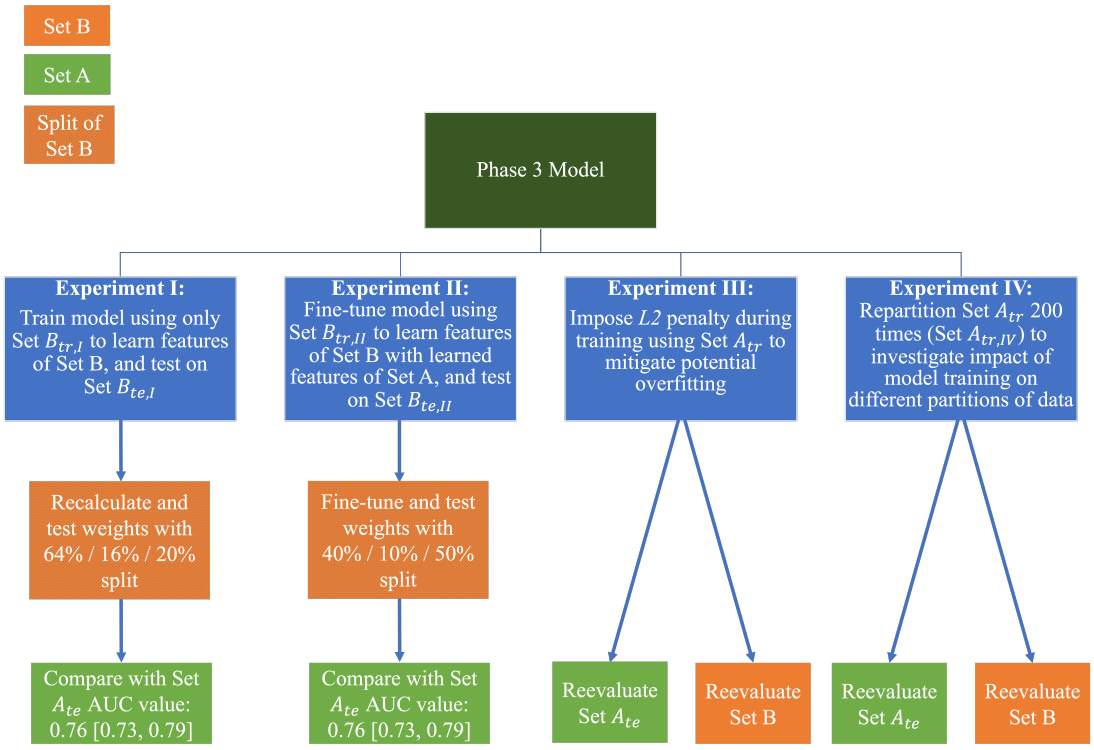


Figure A.2: Summary of the four experiments conducted in this study.

A.3 Results

A.3.1 Experiment I: Recalculating phase 3 weights

After splitting Set B into a 64% training set, 16% validation set, and 20% test set while maintaining the COVID prevalence, the phase 3 weights were recalculated on Set $B_{tr,I}$ and a new AUC value of 0.61 [0.56, 0.66] was obtained on Set $B_{te,I}$ in the task of distinguishing COVID+/- when evaluating the cropped standard CXRs. This value is a significant decrease from 0.67 [0.65, 0.70] ($p = 0.029$), which was obtained when applying the original model to the entirety of Set B. Further, this value was significantly lower than the initial Set A_{te} AUC value of 0.76 [0.73, 0.79] ($p < 0.001$); though, Set $B_{tr,I}$ resulted in fewer images used for training (N = 4,714) when compared with Set A_{tr} (N = 7,888), which could explain the substantial decrease in the AUC values after recalculating the phase 3 weights using Set B.

A.3.2 Experiment II: Fine-tuning phase 3 weights

After fine-tuning the phase 3 weights by splitting the cropped standard CXR images of Set B into 40% training, 10% validation, and 50% testing while maintaining COVID prevalence, the AUC value calculated using Set $B_{te,II}$ slightly improved to 0.70 [0.66, 0.73] but was not significantly different from Set B without fine-tuning the phase 3 weights (AUC = 0.67, $p = 0.27$); though, the 0.70 value was still significantly different from Set A_{te} (AUC = 0.76, $p = 0.007$). A summary of the previous two comparisons is presented in Figure A.3 below.

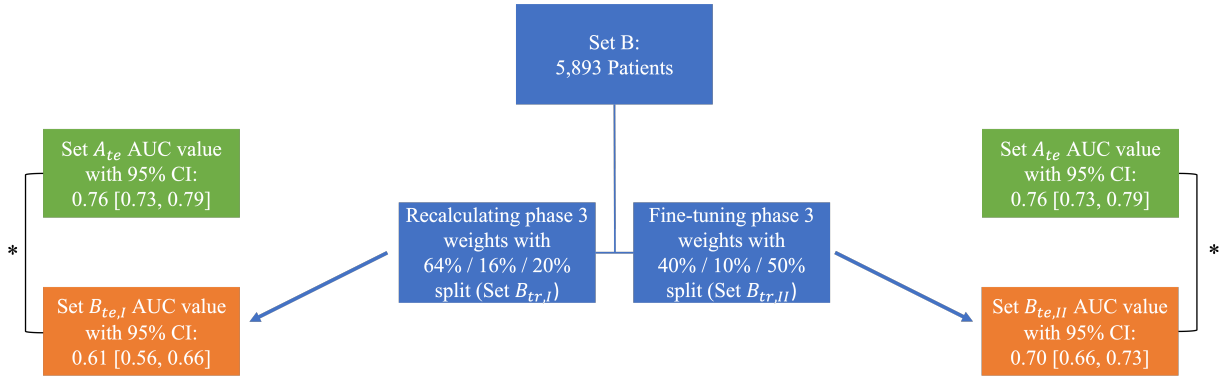


Figure A.3: Summary of the results when recalculating (left) and fine-tuning (right) the phase 3 weights of the model. AUC values calculated in the task of distinguishing COVID+/- CXRs were significantly lower when comparing the partitioned Set B (Set $B_{te,I}$ and Set $B_{te,II}$) results with Set A_{te}, denoted by the asterisks. Green and orange boxes indicate results on Set A and Set B, respectively.

A.3.3 Experiment III: L2 regularization

Regularization did not mitigate model overfitting as the AUC values obtained with the regularized model failed to achieve a significantly higher AUC value than the corresponding AUC values prior to regularization for both Set A_{te} (0.76 [0.72, 0.79]) and Set B (0.68 [0.66, 0.70]).

A.3.4 Experiment IV: Recalculating phase 3 weights after repartitioning

Retraining the model with the Set $A_{tr,IV}$ repartitions using the cropped standard CXR images resulted in an average AUC value of 0.71 $\pm$ 0.013 on Set A_{te} and an average AUC value of 0.66 $\pm$ 0.009 on Set B. There was a Gaussian-like distribution of AUC values for Set B (skew of -0.14) but a slight left-tailed distribution (skew of -0.46) for Set A_{te}, as shown in Figure A.4. There was also a significantly larger variance of AUC values for Set A_{te} than for Set B (F-test, $p < 0.01$), demonstrating the larger impact different training partitions had on Set A_{te} than Set B.

The lowest AUC value achieved on Set A_{te} during the 200 partitions was 0.66 [0.62, 0.69]. Interestingly, the initial AUC value of Set B (0.67 [0.65, 0.70]) set was no longer significantly less than the AUC value obtained with this repartition of Set $A_{tr,IV}$ ($p = 0.46$). Further, this lowest AUC value was significantly less than the initial Set A_{te} AUC value of 0.76 [0.73, 0.79] ($p < 0.001$). The highest value achieved on Set A_{te} during the repartitions was 0.73 [0.70, 0.76], lower than the initial AUC value of 0.76, but this difference just failed to achieve statistical significance ($p = 0.069$).

The lowest AUC value achieved on Set A_{te} from the aforementioned Set $A_{tr,IV}$ repartition (0.66 [0.62, 0.69]) was compared to its corresponding Set B AUC value (0.64 [0.62, 0.66]) on the exact same repartition and failed to achieve a significant difference ($p = 0.43$). Though, the highest AUC value achieved on Set A_{te} from the repartitions (0.73 [0.70, 0.76]) was significantly different from its corresponding Set B AUC value (0.65 [0.63, 0.68]) ($p < 0.001$).

Distributions of AUC values resulting from the 200 partitions are displayed below in Figure A.4 and a summary of the previous two analyses is presented in Figure A.5. There was a significant difference between the distributions of AUC values of Set A_{te} and Set B (Wilcoxon rank-sum test, $p < 0.001$).

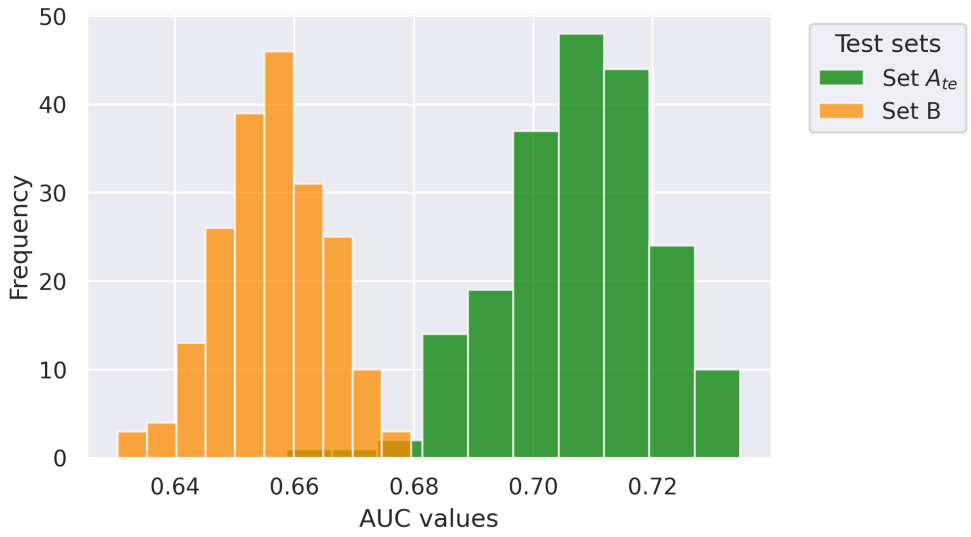


Figure A.4: Distributions of the AUC values obtained when repartitioning Set $A_{tr,IV}$ 200 times and evaluating it on the test set of Set A, Set A_{te}, and the entirety of Set B.

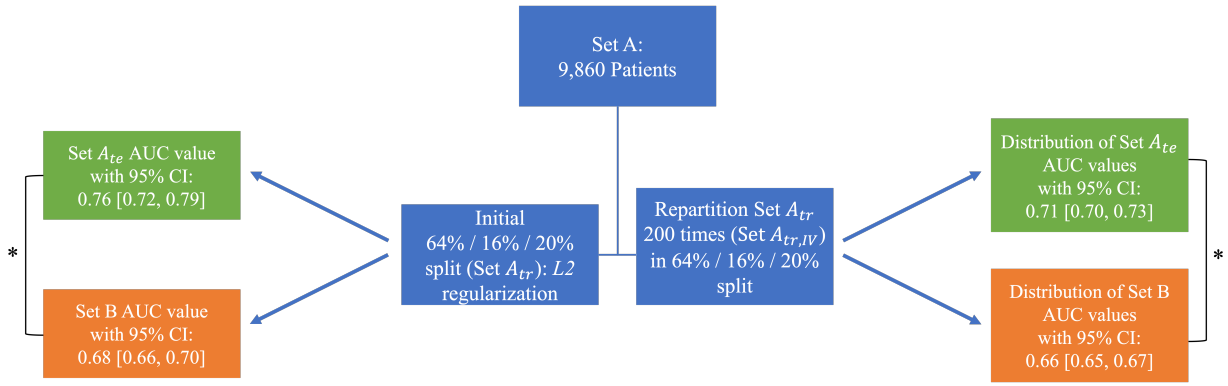


Figure A.5: Summary of the results when implementing $L2$ regularization (left) and repartitioning Set A_{tr} 200 times (right). The AUC value of Set B was significantly lower than Set A_{te} for the $L2$ regularization, denoted by the asterisk. The distributions of the Set A_{te} AUC values and Set B AUC values obtained using the repartitioned Set $A_{tr,IV}$ achieved a significant difference, denoted by the asterisk. Green and orange boxes indicate results on Set A and Set B, respectively.

A.3.5 MIDRC Grand Challenge

The MIDRC COVIDx Grand Challenge comprised portable CXRs and was split in three stages: training, validation, and test. The original DL model trained on Set A_{tr} was applied to the 197 cases in the challenge validation set and achieved an AUC value of 0.60. For reference, the highest-performing AUC value during the validation stage was reported to be 0.66. The winning model, i.e., the model that achieved the highest AUC value during the test stage, obtained a value of 0.70 [159]. Table A.3 displays the AUC values of all the different experiments conducted in this work.

Table A.3: Summary of the main strategies or applications and their corresponding AUC values.

Experiment	Strategy or application	AUC values
I	Recalculating phase 3 weights	Set $B_{te,I}$: 0.61 [0.56, 0.66]
II	Fine-tuning phase 3 weights	Set $B_{te,II}$: 0.70 [0.66, 0.73]
III	$L2$ regularization applied during phase 3	Set A_{te}: 0.76 [0.72, 0.79] Set B: 0.68 [0.66, 0.70]
IV	200 repartitions and recalculating phase 3 weights	Mean of Set A_{te}: 0.71 [0.70, 0.73] Mean of Set B: 0.66 [0.65, 0.67]
—	MIDRC Grand Challenge	0.60 (95% CI was not provided)

A.4 Discussion

The motivation for this study was to examine the potential reasons behind the significant decrease of model performance between the initially partitioned Set A and Set B, which were both acquired at the same institution. Prior work discussed in Chapter 4 and Shenouda et al. [126] extensively investigated this discrepancy in performance of the model between the test sets as it explored impact of age and sex, immunization status, COVID severity, type of imaging equipment, and date matching of image acquisition [126]. None of these studies in

Chapter 4, however, were able to explain the drop in performance. Therefore, this current work examined the impact data partitioning and model retraining have on performance and model generalizability.

Recalculating the phase 3 weights using Set $B_{tr,I}$ in Experiment I failed to improve performance of the model, perhaps due to the smaller number of cases on which the model trained compared with Set A_{tr}, i.e., the original model trained on 6,310 patients from Set A, while the recalculated phase 3 weights were trained on 3,771 patients from Set B. Fine-tuning in Experiment II was performed to incorporate images from Set B in the training scheme (Set $B_{tr,II}$) in an attempt to improve model generalizability. Fine-tuning slightly improved the AUC value from 0.67 on Set B to 0.70 on Set $B_{te,II}$ for the cropped standard CXRs, but that value remained significantly lower than that of the initial Set A_{te} AUC value (0.76). The architecture of the model itself was altered in Experiment III in an attempt to create a more generalizable model. Specifically, $L2$ regularization was implemented to control for overfitting [160], though the regularization had a negligible impact on the performance of the model. Lastly, when applying the original model to a completely external dataset during the MIDRC Grand Challenge, i.e., not from the same institution as the CXRs used for training and testing the model, it yielded an AUC value of 0.60. However, considering the top-performing model achieved an AUC value of 0.66 during validation, the model of this study did not substantially underperform.

AUC values for Set A_{te} had a larger span (range: 0.66–0.73) than those of Set B (range: 0.63–0.68) when repartitioning Set A_{tr} 200 times during Experiment IV. Significant differences were achieved when comparing the highest AUC value calculated on Set A_{te} with its corresponding AUC value on Set B. Further, the highest Set A_{te} AUC value (0.73) failed to achieve a significant difference from the initial 0.76 AUC value, although it was lower. When analyzing the lowest AUC values, the difference between Set A_{te} and its corresponding Set B failed to achieve a significant difference. Differences between the lowest AUC value of Set

A_{te} and the initial 0.67 AUC value of Set B also failed to achieve a significant difference. In other words, these values demonstrate that different partitions of the same dataset will yield significantly different results, returning variable performance. Therefore, while none of the patient demographics and clinical factors of the former study in Chapter 4 could explain the decreased performance of the original model, the repartitioned results here indicate that a favorable, random partition may have been the reason for discrepancy in performance. This work also emphasizes the "black box" nature of DL, as no discernible, real-world characteristic could explain the discrepancy of the model outputs. Instead, multiple repartitionings of the dataset demonstrated the large range of AUC values calculated, and consequently, the breadth of model performance and lack of generalizability. Additionally, these results suggest that DL studies should report on model performance across multiple repartitions of the data, as that would provide a more reliable assessment of the model. Overall, this work is novel as it provides an exhaustive and in-depth analysis investigating different training strategies to explain the decreased model performance when evaluating datasets that were acquired from the same institution.

Future work will explore further the creation of a generalizable model. This will include various regularization and augmentation methods. For example, test-time augmentation (TTA) could be employed by creating multiple augmented versions of the images in the test set. The model then makes predictions on each of these augmented versions, returning an ensemble of predictions, which can then be averaged. Specifically, test entropy minimization can be used to perform the TTA, as the minimization has been shown to reduce generalization error for image classification on corrupted ImageNet, ImageNet-C, and CIFAR-10/100 datasets [161]. In addition, analyses in finding an optimal ratio of the data split into training, validation, and test sets will be conducted. Multiple studies [162, 163, 164, 165] have recommended a variety of splits, ranging from a 25% to a 50% split in creating the test set. Therefore, an optimal ratio will be explored to ensure the model is not overfit, which

may result in improved generalizability. Lastly, an analysis of patient-based performance will be conducted. For instance, subset analyses (i.e., age or sex) can be performed on Set A_{te}, and the classifier outputs, which varied with the different training repartitions, can be studied using a metric such as sureness introduced by Whitney et al. [166] that evaluates the repeatability of the outputs. The metric can be used across different categories and across the two test sets, Set A_{te} and Set B.

A.5 Conclusion

This study examined a model trained to classify COVID-19 status based on patient CXRs and investigated the discrepancy in performance when the model was applied to two separate datasets acquired from the same institution, Set A and Set B. The model yielded significantly different AUC values between the initial test set Set A_{te} (0.76) and newer Set B (0.67). Methods and modifications of model architecture, model retraining, and model fine-tuning were all performed in an attempt to explain the lower AUC value. The exploration of data partitioning was able to provide an explanation for the decreased performance between the datasets, as it underscored the variability introduced by different partitions.

Overall, this work contributes to the methods of explainable AI, as it attempted to interpret the results of the DL algorithm used to classify COVID-19 status. These findings emphasize the need for continued research in improving model training, fine-tuning, and augmentation to address model generalizability before deployment in the clinic.

REFERENCES

[1] H. Kautz. The Third AI Summer: AAAI Robert S. Engelmore Memorial Lecture. *AI Magazine*, 43(1):105–125, 2022. doi:10.1002/aaai.12036. URL https://doi.org/10.1002/aaai.12036.

[2] F. Chollet. *Deep Learning with Python*. Manning Publications, Shelter Island, NY, 2017.

[3] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. The MIT Press, Cambridge, MA, 2016.

[4] Stanford CS Department. CS231n Convolutional Neural Networks for Computer Vision, 2023. URL https://cs231n.github.io/neural-networks-2/.

[5] Stanford CS Department. CS231n Convolutional Neural Networks for Computer Vision, 2023. URL https://cs231n.github.io/convolutional-networks/.

[6] C. Guo, G. Pleiss, et al. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 1321–1330. JMLR.org, 2017. URL https://dl.acm.org/doi/10.5555/3305381.3305518.

[7] Z. Ding, X. Han, et al. Local temperature scaling for probability calibration. *arXiv preprint*, 2021. doi:10.48550/arXiv.2008.05105. URL https://doi.org/10.48550/arXiv.2008.05105.

[8] M. Bojarski, D.D. Testa, et al. End to end learning for self-driving cars. *arXiv preprint*, 2016. doi:10.48550/arXiv.1604.07316. URL https://doi.org/10.48550/arXiv.1604.07316.

[9] M. Havaei, A. Davy, et al. Brain tumor segmentation with deep neural networks. *Med Image Anal*, 35:18–31, 2017. doi:10.1016/j.media.2016.05.004. URL https://doi.org/10.1016/j.media.2016.05.004.

[10] Z. Li, K. Kamnitsas, and B. Glocker. Overfitting of neural nets under class imbalance: Analysis and improvements for segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 402–410. Springer, 2019. doi:10.1007/978-3-030-32248-9_45. URL https://doi.org/10.1007/978-3-030-32248-9_45.

[11] A. Garcia-Garcia, S. Orts-Escolano, et al. A review on deep learning techniques applied to semantic segmentation. *arXiv preprint*, 2017. doi:10.48550/arXiv.1704.06857. URL https://doi.org/10.48550/arXiv.1704.06857.

[12] R. Jena and S.P. Awate. A Bayesian neural net to segment images with uncertainty estimates and good calibration. In *International Conference on Information Processing in Medical Imaging*, pages 3–15. Springer, 2019.

[13] A.S. Mozafari, H.S. Gomes, et al. Attended temperature scaling: A practical approach for calibrating deep neural networks. *arXiv preprint*, 2019. doi:10.48550/arXiv.1810.11586. URL https://doi.org/10.48550/arXiv.1810.11586.

[14] S.A. Balanya, J. Maroñas, and D. Ramosa. Adaptive temperature scaling for robust calibration of deep neural networks. *arXiv preprint*, 2022. doi:10.48550/arXiv.2208.00461. URL https://doi.org/10.48550/arXiv.2208.00461.

[15] D. Wang, B. Wong, and L. Wang. On calibrating semantic segmentation models: Analyses and an algorithm. *arXiv preprint*, 2023. doi:10.48550/arXiv.2212.12053. URL https://doi.org/10.48550/arXiv.2212.12053.

[16] J. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in large margin classifiers*. The MIT Press, Cambridge, USA, 1999.

[17] Y. Gal and Z. Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. *arXiv preprint*, 2016. doi:10.48550/arXiv.1506.02142. URL https://doi.org/10.48550/arXiv.1506.02142.

[18] T.M. Elsheikh, S.L. Asa, et al. Interobserver and intraobserver variation among experts in the diagnosis of thyroid follicular lesions with borderline nuclear features of papillary carcinoma. *Am J Clin Pathol*, 130(5):736–744, November 2008. doi:10.1309/AJCPKP2QUVN4RCCP. URL https://doi.org/10.1309/AJCPKP2QUVN4RCCP.

[19] Q. Li and K. Doi. Analysis and minimization of overtraining effect in rule-based classifiers for computer-aided diagnosis. *Med Phys*, 33(2):320–328, 2006. doi:10.1118/1.1999126. URL https://doi.org/10.1118/1.1999126.

[20] K. Doi, M. L. Giger, et al. Computer aided diagnosis of breast cancer on mammograms. *Breast Cancer*, 4(4):228–233, 1997. doi:10.1007/BF02966511. URL https://doi.org/10.1007/BF02966511.

[21] J.N. Wolfe. Breast patterns as an index of risk for developing breast cancer. *AJR Am J Roentgenol*, 126(6):1130–1137, 1976. doi:10.2214/ajr.126.6.1130. URL https://doi.org/10.2214/ajr.126.6.1130.

[22] E. Warner, G. Lockwood, et al. The risk of breast cancer associated with mammographic parenchymal patterns: a meta-analysis of the published literature to examine the effect of the method of classification. *Cancer Detect Prev*, 16(1):67–72, 1992.

[23] M.L. Giger, H.P. Chan, and J. Boone. Anniversary paper: History and status of CAD and quantitative image analysis: The role of medical physics and AAPM. *Med Phys*, 35(12):5799–5820, 2008. doi:10.1118/1.3013555. URL https://doi.org/10.1118/1.3013555.

[24] Z. Huo, M.L. Giger, et al. Computerized analysis of mammographic parenchymal patterns for breast cancer risk assessment: feature selection. *Med Phys*, 27(1):4–12, 2000. doi:10.1118/1.598851. URL https://doi.org/10.1118/1.598851.

[25] P. W. Maragos. Fractal signal analysis using mathematical morphology. In P. W. Hawkes, editor, *Advances in Electronics and Electron Physics*, volume 88, pages 199–246. Academic Press: Harcourt Brace and Company, NY, NY, 1994.

[26] K.I. Laws. Rapid Texture Identification. In Thomas F. Wiener, editor, *Image Processing for Missile Guidance*, volume 0238, pages 376–381. International Society for Optics and Photonics, SPIE, 1980. doi:10.1117/12.959169. URL https://doi.org/10.1117/12.959169.

[27] H. Li, M. L. Giger, et al. Power spectral analysis of mammographic parenchymal patterns for breast cancer risk assessment. *J Digit Imaging*, 21(2):145–152, 2008. doi:10.1007/s10278-007-9093-9. URL https://doi.org/10.1007/s10278-007-9093-9.

[28] J.J. Foy, K.R. Robinson, et al. Variation in algorithm implementation across radiomics software. *J Med Imaging*, 5, 2018. doi:10.1117/1.JMI.5.4.044505. URL https://doi.org/10.1117/1.JMI.5.4.044505.

[29] H. Li, M. L. Giger, et al. Fractal analysis of mammographic parenchymal patterns in breast cancer risk assessment. *Acad Radiol*, 14(5):513–521, 2007. doi:10.1016/j.acra.2007.02.003. URL https://doi.org/10.1016/j.acra.2007.02.003.

[30] R. F. Chang, W. J. Wu, et al. Improvement in breast tumor discrimination by support vector machines and speckle-emphasis texture analysis. *Ultrasound Med Biol*, 29(5):679–686, 2003. doi:10.1016/s0301-5629(02)00788-3. URL https://doi.org/10.1016/s0301-5629(02)00788-3.

[31] R. Simon, M. D. Radmacher, et al. Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *J Natl Cancer Inst*, 95(1):14–18, 2003. doi:10.1093/jnci/95.1.14. URL https://doi.org/10.1093/jnci/95.1.14.

[32] P. Lambin, E. Rios-Velazquez, et al. Radiomics: Extracting more information from medical images using advanced feature analysis. *Eur J Cancer*, 48(4):441–446, Mar 2012. doi:10.1016/j.ejca.2011.11.036. URL https://doi.org/10.1016/j.ejca.2011.11.036. Epub 2012 Jan 16.

[33] N. Petrick. Pre- and post-market evaluation of autonomous ai/ml: Lessons learned from prior cad devices, 2020. URL https://www.fda.gov/media/135712/download.

[34] M.L. Giger. Machine learning in medical imaging. *J Am Coll Radiol*, 15:512–520, 2018. doi:10.1016/j.jacr.2017.12.028. URL https://doi.org/10.1016/j.jacr.2017.12.028.

[35] I.G.M.L. el Naqa, M.A. Haider, et al. Artificial intelligence: reshaping the practice of radiological sciences in the 21st century. *Br J Radiol*, 93(1106):20190855, 2020. doi:10.1259/bjr.20190855. URL https://doi.org/10.1259/bjr.20190855.

[36] S.-C.B. Lo, H.-P. Chan, et al. Artificial convolution neural network for medical image pattern recognition. *Neural Networks*, 8(7-8):1201–1214, 1995. doi:10.1016/0893-6080(95)00061-5. URL https://doi.org/10.1016/0893-6080(95)00061-5.

[37] W. Zhang, K. Doi, et al. Computerized detection of clustered microcalcifications in digital mammograms using a shift-invariant artificial neural network. *Med Phys*, 21: 517–524, 1994. doi:10.1118/1.597177. URL https://doi.org/10.1118/1.597177.

[38] A.S. Lundervold and A. Lundervold. An overview of deep learning in medical imaging focusing on MRI. *Z Med Phys*, 29:102–127, 2019. doi:10.1016/j.zemedi.2018.11.002. URL https://doi.org/10.1016/j.zemedi.2018.11.002.

[39] S.S. Yadav and S.M. Jadhav. Deep convolutional neural network based medical image classification for disease diagnosis. *J Big Data*, 6, 2019. doi:10.1186/s40537-019-0276-2. URL https://doi.org/10.1186/s40537-019-0276-2.

[40] R. Yamashita, M. Nishio, et al. Convolutional neural networks: an overview and application in radiology. *Insights Imaging*, 9:611–629, 2018. doi:10.1007/s13244-018-0639-9. URL https://doi.org/10.1007/s13244-018-0639-9.

[41] R. Yang and Y. Yu. Artificial convolutional neural network in object detection and semantic segmentation for medical imaging analysis. *Front Oncol*, 11, 2021. doi:10.3389/fonc.2021.638182. URL https://doi.org/10.3389/fonc.2021.638182.

[42] E. Gudmundsson, C.M. Straus, and S.G. Armato III. Deep convolutional neural networks for the automated segmentation of malignant pleural mesothelioma on computed tomography scans. *J Med Imaging*, 5:034503, 2018. doi:10.1117/1.jmi.5.3.034503. URL https://doi.org/10.1117/1.jmi.5.3.034503.

[43] E. Gudmundsson, C.M. Straus, et al. Deep learning-based segmentation of malignant pleural mesothelioma tumor on computed tomography scans: application to scans demonstrating pleural effusion. *J Med Imaging*, 7:012705, 2020. doi:10.1117/1.jmi.7.1.012705. URL https://doi.org/10.1117/1.jmi.7.1.012705.

[44] Q. Hu, K. Drukker, and M.L. Giger. Role of standard and soft tissue chest radiography images in deep-learning-based early diagnosis of COVID-19. *J Med Imaging (Bellingham)*, 8, 2021. doi:10.1117/1.JMI.8.S1.014503. URL https://doi.org/10.1117/1.JMI.8.S1.014503.

[45] M. Ray and H.L. Kindler. Malignant pleural mesothelioma: An update on biomarkers and treatment. *Chest*, 136:888–896, 2009. doi:10.1378/chest.08-2665. URL https://doi.org/10.1378/chest.08-2665.

[46] N.P. Campbell and H.L. Kindler. Update on malignant pleural mesothelioma. *Semin Respir Crit Care Med*, 32:102–110, 2011. doi:10.1055/s-0031-1272874. URL https://doi.org/10.1055/s-0031-1272874.

[47] A.C. Bibby, S. Tsim, et al. Malignant pleural mesothelioma: An update on investigation, diagnosis and treatment. *Eur Respir Rev*, 25:472–486, 2016. doi:10.1183/16000617.0063-2016. URL https://doi.org/10.1183/16000617.0063-2016.

[48] S.I. Katz, C.M. Straus, et al. Considerations for imaging of malignant pleural mesothelioma: A consensus statement from the international mesothelioma interest group. *J Thorac Oncol*, 18(3):278–298, 2023. ISSN 1556-0864. doi:10.1016/j.jtho.2022.11.018. URL https://doi.org/10.1016/j.jtho.2022.11.018.

[49] A.N. Husain, T.V. Colby, et al. Guidelines for pathologic diagnosis of malignant mesothelioma: 2017 update of the consensus statement from the international mesothelioma interest group. *Arch Pathol Lab Med*, 142:89–108, 2018. doi:10.5858/arpa.2017-0124-RA. URL https://doi.org/10.5858/arpa.2017-0124-RA.

[50] L. Righi, E. Duregon, et al. BRCA1-associated protein 1 (BAP1) immunohistochemical expression as a diagnostic tool in malignant pleural mesothelioma classification: A large retrospective study. *J Thorac Oncol*, 11:2006–2017, 2016. doi:10.1016/j.jtho.2016.06.020. URL https://doi.org/10.1016/j.jtho.2016.06.020.

[51] H.L. Kindler, N. Ismaila, et al. Treatment of malignant pleural mesothelioma: American society of clinical oncology clinical practice guideline. *JCO*, 36:1343–1373, 2018. doi:10.1200/JCO.2017.76.6394. URL https://doi.org/10.1200/JCO.2017.76.6394.

[52] L. Cardinale, F. Ardissone, et al. Diagnostic imaging and workup of malignant pleural mesothelioma. *Acta Biomed*, 88:134–142, 2017. doi:10.23750/ABM.V88I2.5558. URL https://doi.org/10.23750/ABM.V88I2.5558.

[53] M.-Y. Ng, E.Y.P. Lee, et al. Imaging profile of the COVID-19 infection: Radiologic findings and literature review. *Radiol Cardiothorac Imaging*, 2(1):e200034, 2020. doi:10.1148/ryct.2020200034. URL https://doi.org/10.1148/ryct.2020200034.

[54] H.Y.F. Wong, H.Y.S. Lam, et al. Frequency and distribution of chest radiographic findings in patients positive for COVID-19. *Radiology*, 296(2), 2020. doi:10.1148/radiol.2020201160. URL https://doi.org/10.1148/radiol.2020201160.

[55] M.V. Gerwen, N. Alpert, et al. Prognostic factors of survival in patients with malignant pleural mesothelioma: An analysis of the national cancer database. *Carcinogenesis*, 40 (4):529–536, 2019. doi:10.1093/carcin/bgz004. URL https://doi.org/10.1093/carcin/bgz004.

[56] S.G. Armato III, K.G. Blyth, et al. Imaging in pleural mesothelioma: A review of the 13th international conference of the international mesothelioma interest group. *Lung Cancer*, 101:48–58, 2016. doi:10.1016/j.lungcan.2016.09.003. URL https://doi.org/10.1016/j.lungcan.2016.09.003.

[57] S.G. Armato III, R.J. Francis, et al. Imaging in pleural mesothelioma: A review of the 14th international conference of the international mesothelioma interest group. *Lung Cancer*, 130:108–114, 2019. doi:10.1016/j.lungcan.2018.11.033. URL https://doi.org/10.1016/j.lungcan.2018.11.033.

[58] S.G. Armato III, A.K. Nowak, et al. Imaging in pleural mesothelioma: A review of the 15th international conference of the international mesothelioma interest group. *Lung Cancer*, 164:76–83, 2022. doi:10.1016/j.lungcan.2021.12.008. URL https://10.1016/j.lungcan.2021.12.008.

[59] M.J. Byrne and A.K. Nowak. Modified recist criteria for assessment of response in malignant pleural mesothelioma. *Ann Oncol*, 15(2):257–260, 2004. doi:10.1093/annonc/mdh059. URL https://doi.org/10.1093/annonc/mdh059.

[60] G.R. Oxnard, S.G. Armato III, and H.L. Kindler. Modeling of mesothelioma growth demonstrates weaknesses of current response criteria. *Lung Cancer*, 52(2):141–148, 2006. doi:10.1016/j.lungcan.2005.12.013. URL https://doi.org/10.1016/j.lungcan.2005.12.013.

[61] S.G. Armato III and A.K. Nowak. Revised modified response evaluation criteria in solid tumors for assessment of response in malignant pleural mesothelioma (version 1.1). *J Thorac Oncol*, 13(7):1012–1021, 2018. doi:10.1016/j.jtho.2018.04.034. URL https://doi.org/10.1016/j.jtho.2018.04.034.

[62] E.A. Eisenhauer, P. Therasse, et al. New response evaluation criteria in solid tumours: Revised recist guideline (version 1.1). *Eur J Cancer*, 45(2):228–247, 2009. doi:10.1016/j.ejca.2008.10.026. URL https://doi.org/10.1016/j.ejca.2008.10.026.

[63] X.J. Xie, S.Y. Liu, et al. Development of unenhanced ct-based imaging signature for BAP1 mutation status prediction in malignant pleural mesothelioma: Consideration of 2d and 3d segmentation. *Lung Cancer*, 157:30–39, 2021. ISSN 0169-5002. doi:10.1016/j.lungcan.2021.04.023. URL https://doi.org/10.1016/j.lungcan.2021.04.023.

[64] H.I. Pass, B.K. Temeck, et al. Preoperative tumor volume is associated with outcome in malignant pleural mesothelioma. *J Thorac Cardiovasc Surg*, 115(2):310–318, 1998. doi:10.1016/s0022-5223(98)70274-0. URL https://doi.org/10.1016/s0022-5223(98)70274-0.

[65] D.J. Murphy and R.R. Gill. Volumetric assessment in malignant pleural mesothelioma. *Ann Transl Med*, 5(11):241–241, 2017. doi:10.21037/atm.2017.05.23. URL https://doi.org/10.21037/atm.2017.05.23.

[66] W.F. Sensakovic, S.G. Armato III, et al. Computerized segmentation and measurement of malignant pleural mesothelioma. *Med Phys*, 38(1):238–244, 2011. doi:10.1118/1.3525836. URL https://doi.org/10.1118/1.3525836.

[67] K. Zormpas-Petridis, N. Tunariu, et al. Accelerating whole-body diffusion-weighted MRI with deep learning-based denoising image filters. *Radiol Artif Intell*, 3(5):e200279, 2021. doi:10.1148/ryai.2021200279. URL https://doi.org/10.1148/ryai.2021200279.

[68] J.R. Naso, A.B. Levine, et al. Deep-learning based classification distinguishes sarcomatoid malignant mesotheliomas from benign spindle cell mesothelial proliferations. *Mod Pathol*, 34(11):2028–2035, 2021. doi:10.1038/s41379-021-00850-6. URL https://doi.org/10.1038/s41379-021-00850-6.

[69] H. Matsuo, K. Kitajima, et al. Prognosis prediction of patients with malignant pleural mesothelioma using conditional variational autoencoder on 3D PET images and clinical data. *Med Phys*, 50(12):7548–7557, 2023. doi:10.1002/mp.16694. URL https://doi.org/10.1002/mp.16694.

[70] A.C. Kidd, O. Anderson, et al. Fully automated volumetric measurement of malignant pleural mesothelioma by deep learning AI: Validation and comparison with modified recist response criteria. *Thorax*, 77(12):1251–1259, 2022. doi:10.1136/thoraxjnl-2021-217808. URL https://doi.org/10.1136/thoraxjnl-2021-217808.

[71] P. Godau, P. Kalinowski, et al. Deployment of image analysis algorithms under prevalence shifts. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, volume 14222 of *Lecture Notes in Computer Science*. Springer, 2023. doi:10.1007/978-3-031-43898-1_38. URL https://doi.org/10.1007/978-3-031-43898-1_38.

[72] M. Shenouda, E. Gudmundsson, et al. Convolutional neural networks for segmentation of pleural mesothelioma: Analysis of probability map thresholds (CALGB 30901, alliance). *JIIM*, 2024. In press.

[73] A.Z. Dudek, X. Wang, et al. Randomized study of maintenance pemetrexed versus observation for treatment of malignant pleural mesothelioma: CALGB 30901. *Clin Lung Cancer*, 21(6), 2020. doi:10.1016/j.cllc.2020.06.025. URL https://doi.org/10.1016/j.cllc.2020.06.025.

[74] O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional networks for biomedical image segmentation. *Lecture Notes in Computer Science*, page 234–241, 2015. doi:10.1007/978-3-319-24574-4_28. URL https://link.springer.com/chapter/10.1007/978-3-319-24574-4_28.

[75] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

[76] A.A. Taha and A. Hanbury. Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool. *BMC Med Imaging*, 15(1), 2015. doi:10.1186/s12880-015-0068-x. URL https://doi.org/10.1186/s12880-015-0068-x.

[77] F.J. Massey. The Kolmogorov-Smirnov test for goodness of fit. *J Am Stat Assoc*, 46 (253):68–78, 1951. doi:10.1080/01621459.1951.10500769. URL https://doi.org/10.1080/01621459.1951.10500769.

[78] J.M. Bland and D.G. Altman. Measuring agreement in method comparison studies. *Stat Methods Med Res*, 8(2):135–160, 1999. doi:10.1177/096228029900800204. URL https://doi.org/10.1177/096228029900800204.

[79] W.F. Sensakovic, A. Starkey, et al. The influence of initial outlines on manual segmentation. *Med Phys*, 37(5):2153–2158, 2010. doi:10.1118/1.3392287. URL https://doi.org/10.1118/1.3392287.

[80] R.J. Gillies, P.E. Kinahan, and H. Hricak. Radiomics: Images are more than pictures, they are data. *Radiology*, 278(2):563–577, 2016. doi:10.1148/radiol.2015151169. URL https://doi.org/10.1148/radiol.2015151169.

[81] W.L. Bi, A. Hosny, et al. Artificial intelligence in cancer imaging: Clinical challenges and applications. *CA Cancer J Clin*, 69(2):127–157, 2019. doi:10.3322/caac.21552. URL https://doi.org/10.3322/caac.21552.

[82] Z. Bodalal, S. Trebeschi, et al. Radiogenomics: Bridging imaging and genomics. *Abdom Radiol (NY)*, 44(6):1960–1984, 2019. doi:10.1007/s00261-019-02028-w. URL https://doi.org/10.1007/s00261-019-02028-w.

[83] E.R. Velazquez, C. Parmar, et al. Somatic mutations drive distinct imaging phenotypes in lung cancer. *Cancer Res*, 77(14):3922–3930, 2017. doi:10.1158/0008-5472.CAN-17-0122. URL https://doi.org/10.1158/0008-5472.CAN-17-0122.

[84] S.S. Yip, J. Kim, et al. Associations between somatic mutations and metabolic imaging phenotypes in non-small cell lung cancer. *J Nucl Med*, 58(4):569–576, 2017. doi:10.2967/jnumed.116.181826. URL https://doi.org/10.2967/jnumed.116.181826.

[85] M. Cigognetti, S. Lonardi, et al. BAP1 (BRCA1-associated protein 1) is a highly specific marker for differentiating mesothelioma from reactive mesothelial proliferations. *Mod Pathol*, 28(8):1043–1057, 2015. doi:10.1038/modpathol.2015.65. URL https://doi.org/10.1038/modpathol.2015.65.

[86] R. Murali, T. Wiesner, and R. A. Scolyer. Tumours associated with BAP1 mutations. *Pathology*, 45(2):116–126, 2013. doi:10.1097/PAT.0b013e32835d0efb. URL https://doi.org/10.1097/PAT.0b013e32835d0efb.

[87] M. Nasu, M. Emi, et al. High incidence of somatic BAP1 alterations in sporadic malignant mesothelioma. *J Thorac Oncol*, 10(4):565–576, 2015. doi:10.1097/JTO.0000000000000471. URL https://doi.org/10.1097/JTO.0000000000000471.

[88] O. D. Mitchell, K. Gilliam, et al. Germline variants incidentally detected via tumor-only genomic profiling of patients with mesothelioma. *JAMA Netw Open*, 6(8): e2327351, Aug 1 2023. doi:10.1001/jamanetworkopen.2023.27351.

[89] M. Carbone, H. I. Pass, et al. Medical and surgical care of patients with mesothelioma and their relatives carrying germline BAP1 mutations. *J Thorac Oncol*, 17(7):873–889, 2022. doi:10.1016/j.jtho.2022.03.014. URL https://doi.org/10.1016/j.jtho.2022.03.014.

[90] F. Baumann, E. Flores, et al. Mesothelioma patients with germline BAP1 mutations have 7-fold improved long-term survival. *Carcinogenesis*, 36(1):76–81, 2015. doi:10.1093/carcin/bgu227. URL https://doi.org/10.1093/carcin/bgu227.

[91] M. B. Daly, T. Pal, et al. Genetic/familial high-risk assessment: Breast, ovarian, and pancreatic, version 2.2021, NCCN clinical practice guidelines in oncology. *J Natl Compr Canc Netw*, 19(1):77–102, 2021. doi:10.6004/jnccn.2021.0001. URL https://doi.org/10.6004/jnccn.2021.0001.

[92] M. Shenouda, A. Shaikh, et al. The use of radiomics on computed tomography scans for differentiation of somatic BAP1 mutation status for patients with pleural mesothelioma. In *Proc SPIE*, volume 12927, Apr. 2024. doi:10.1117/12.3000085. URL https://doi.org/10.1117/12.3000085.

[93] M. Carbone, H. Yang, et al. BAP1 and cancer. *Nat Rev Cancer*, 13(3):153–159, Mar 2013. doi:10.1038/nrc3459. URL https://doi.org/10.1038/nrc3459.

[94] J.R. Testa, M. Cheung, et al. Germline BAP1 mutations predispose to malignant mesothelioma. *Nat Genet*, 43(10):1022–1025, Aug 2011. doi:10.1038/ng.912. URL https://doi.org/10.1038/ng.912.

[95] M.G. Zauderer, M. Bott, et al. Clinical characteristics of patients with malignant pleural mesothelioma harboring somatic BAP1 mutations. *J Thorac Oncol*, 8(11): 1430–1433, Nov 2013. doi:10.1097/JTO.0b013e31829e7ef9. URL https://doi.org/10.1097/JTO.0b013e31829e7ef9.

[96] S. Kadri, B. C. Long, et al. Clinical validation of a next-generation sequencing genomic oncology panel via cross-platform benchmarking against established amplicon sequencing assays. *J Mol Diagn*, 19(1):43–56, 2017. doi:10.1016/j.jmoldx.2016.07.012. URL https://doi.org/10.1016/j.jmoldx.2016.07.012.

[97] S.G. Armato III, G. McLennan, et al. The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): a completed reference database of

lung nodules on CT scans. *Med Phys*, 38(2):915–931, 2011. doi:10.1118/1.3528204. URL https://doi.org/10.1118/1.3528204.

[98] S.G. Armato, N.P. Gruszauskas, et al. Research imaging in an academic medical center. *Acad Radiol*, 19(6):762–771, 2012. ISSN 1076-6332. doi:10.1016/j.acra.2012.02.002. URL https://doi.org/10.1016/j.acra.2012.02.002.

[99] N.P. Gruszauskas and S.G. Armato. Critical challenges to the management of clinical trial imaging: Recommendations for the conduct of imaging at investigational sites. *Acad Radiol*, 27(2):300–306, 2020. ISSN 1076-6332. doi:10.1016/j.acra.2019.04.003. URL https://doi.org/10.1016/j.acra.2019.04.003.

[100] L.R. Dice. Measures of the amount of ecologic association between species. *Ecology*, 26 (3):297–302, 1945. doi:10.2307/1932409. URL https://doi.org/10.2307/1932409.

[101] L. Escudero Sanchez, L. Rundo, et al. Robustness of radiomic features in CT images with different slice thickness, comparing liver tumour and muscle. *Sci Rep*, 11(1):8262, 2021. doi:10.1038/s41598-021-87598-w. URL https://doi.org/10.1038/s41598-021-87598-w.

[102] M. Shafiq-Ul-Hassan, G.G. Zhang, et al. Intrinsic dependencies of CT radiomic features on voxel size and number of gray levels. *Med Phys*, 44(3):1050–1062, 2017. doi:10.1002/mp.12123. URL https://doi.org/10.1002/mp.12123.

[103] A. Zwanenburg, M. Vallières, et al. The image biomarker standardization initiative: Standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology*, 295(2):328–338, 2020. doi:10.1148/radiol.2020191145. URL https://doi.org/10.1148/radiol.2020191145.

[104] J.J.M. van Griethuysen, A. Fedorov, et al. Computational radiomics system to decode the radiographic phenotype. *Cancer Res*, 77(21):e104–e107, 2017. doi:10.1158/0008-5472.CAN-17-0339. URL https://doi.org/10.1158/0008-5472.CAN-17-0339.

[105] G.E.A.P.A. Batista, A.L.C. Bazzan, and M.C. Monard. Balancing training data for automated annotation of keywords: a case study. In *WOB*, 2003. URL https://api.semanticscholar.org/CorpusID:1579194.

[106] A. Niculescu-Mizil and R. Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning*, ICML '05, page 625–632, New York, NY, USA, 2005. Association for Computing Machinery. ISBN 1595931805. doi:10.1145/1102351.1102430. URL https://doi.org/10.1145/1102351.1102430.

[107] A. K. Menon, X. J. Jiang, et al. Predicting accurate probabilities with a ranking loss. In *Proc Int Conf Mach Learn*, pages 703–710, 2012.

[108] E.R. DeLong, D.M. DeLong, and D.L. Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, 44(3):837, 1988. doi:10.2307/2531595. URL https://doi.org/10.2307/2531595.

[109] C.E. Metz and X. Pan. "Proper" binormal ROC curves: Theory and maximum-likelihood estimation. *J Math Psychol*, 43(1):1–33, 1999. doi:10.1006/jmps.1998.1218. URL https://doi.org/10.1006/jmps.1998.1218.

[110] J.E. van Timmeren R.T.H.M. Larue et al. Influence of gray level discretization on radiomic feature stability for different CT scanners, tube currents and slice thicknesses: a comprehensive phantom study. *Acta Oncol*, 56(11):1544–1553, 2017. doi:10.1080/0284186X.2017.1351624. URL https://doi.org/10.1080/0284186X.2017.1351624.

[111] M. Shafiq-ul Hassan, K. Latifi, et al. Voxel size and gray level normalization of CT radiomic features in lung cancer. *Sci Rep*, 8:10545, 2018. doi:10.1038/s41598-018-28895-9. URL https://doi.org/10.1038/s41598-018-28895-9.

[112] B. Efron. Estimating the error rate of a prediction rule: Improvement on cross-validation. *J Am Stat Assoc*, 78(382):316–331, 1983. doi:10.2307/2288636. URL https://doi.org/10.2307/2288636.

[113] Y. Zhang and Y. Yang. Cross-validation for selecting a model selection procedure. *J Econom*, 187(1):95–112, 2015. doi:10.1016/j.jeconom.2015.02.021.

[114] Y. Bengio and Y. Grandvalet. No unbiased estimator of the variance of k-fold cross-validation. *J Mach Learn Res*, 5:1089–1105, dec 2004.

[115] A. Shaikh, I. Deutsch, et al. Assessing radiomic feature robustness using agreement over image perturbation. In *Proc SPIE*, volume 12927, Apr. 2024. doi:10.1117/12.3006291. URL https://doi.org/10.1117/12.3006291.

[116] A. Zwanenburg, S. Leger, et al. Assessing robustness of radiomic features by image perturbation. *Sci Rep*, 9:614, 2019. doi:10.1038/s41598-018-36938-4. URL https://doi.org/10.1038/s41598-018-36938-4.

[117] K. Murphy, H. Smits, et al. COVID-19 on chest radiographs: A multireader evaluation of an artificial intelligence system. *Radiology*, 296(3), 2020. doi:10.1148/radiol.2020201874. URL https://doi.org/10.1148/radiol.2020201874.

[118] R. Zhang, X. Tie, et al. Diagnosis of coronavirus disease 2019 pneumonia by using chest radiography: Value of artificial intelligence. *Radiology*, 298(2), 2021. doi:10.1148/radiol.2020202944. URL https://doi.org/10.1148/radiol.2020202944.

[119] W.H. Chiu, V. Vardhanabhuti, et al. Detection of COVID-19 using deep learning algorithms on chest radiographs. *J Thorac Imaging*, 6, 2020. doi:10.1097/rti.0000000000000559. URL https://doi.org/10.1097/rti.0000000000000559.

[120] R.M. Wehbe, J. Sheng, et al. DeepCOVID-XR: An artificial intelligence algorithm to detect COVID-19 on chest radiographs trained and tested on a large U.S. clinical data set. *Radiology*, 299(1), 2021. doi:10.1148/radiol.2020203511. URL https://doi.org/10.1148/radiol.2020203511.

[121] J.C. Yao, T. Wang, et al. AI detection of mild COVID-19 pneumonia from chest CT scans. *Eur Radiol*, 31(9):7192–7201, 2021. doi:10.1007/s00330-021-07797-x. URL https://doi.org/10.1007/s00330-021-07797-x.

[122] K. Zhang, X. Liu, et al. Clinically applicable AI system for accurate diagnosis, quantitative measurements, and prognosis of COVID-19 pneumonia using computed tomography. *Cell*, 181(6), 2020. doi:10.1016/j.cell.2020.04.045. URL https://10.1016/j.cell.2020.04.045.

[123] M. Roberts, D. Driggs, et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nat Mach Intell*, 3:199–217, 2021. doi:10.1038/s42256-021-00307-0. URL https://doi.org/10.1038/s42256-021-00307-0.

[124] M.B.A. McDermott, S. Wang, et al. Reproducibility in machine learning for health research: Still a ways to go. *Sci Trans Med*, 13, 2021. doi:10.1126/scitranslmed.abb1655. URL https://doi.org/10.1126/scitranslmed.abb1655.

[125] J. Yang, A.A.S. Soltan, and D.A. Clifton. Machine learning generalizability across healthcare settings: insights from multi-site COVID-19 screening. *npj Digit Med*, 5, 2022. doi:10.1038/s41746-022-00614-9. URL https://doi.org/10.1038/s41746-022-00614-9.

[126] M. Shenouda, I. Flerlage, et al. Assessment of a deep learning model for COVID-19 classification on chest radiographs: a comparison across image acquisition techniques and clinical factors. *J Med Imag (Bellingham)*, 10(6):064504, 2023. doi:10.1117/1.JMI.10.6.064504. URL https://doi.org/10.1117/1.JMI.10.6.064504.

[127] M. Shenouda, A. Kaveti, et al. Assessing robustness of a deep learning model for covid-19 classification on chest radiographs. In *Proc SPIE*, volume 12456, Apr. 2023. doi:10.1117/12.2652106. URL https://doi.org/10.1117/12.2652106.

[128] S. Motamed, P. Rogalla, and F. Khalvati. RANDGAN: Randomized generative adversarial network for detection of COVID-19 in chest x-ray. *Sci Rep*, 11(1), 2021. doi:10.1038/s41598-021-87994-2. URL https://doi.org/10.1038/s41598-021-87994-2.

[129] S. Jaeger, S. Candemir, et al. Two public chest x-ray datasets for computer-aided screening of pulmonary diseases. *Quant Imaging Med Surg*, 4(6):475–477, 2014. doi:10.3978/j.issn.2223-4292.2014.11.20. URL https://doi.org/10.3978/j.issn.2223-4292.2014.11.20.

[130] L. Wang, Z.Q. Lin, and A. Wong. COVID-net: A tailored deep convolutional neural network design for detection of COVID-19 cases from chest x-ray images. *Sci Rep*, 10 (1), 2020. doi:10.1038/s41598-020-76550-z. URL https://doi.org/10.1038/s41598-020-76550-z.

[131] G. Huang, Z. Liu, et al. Densely connected convolutional networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. doi:10.1109/cvpr.2017.243. URL https://doi.org/10.1109/cvpr.2017.243.

[132] P. Rajpurkar, J. Irvin, et al. CheXNet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint*, 2017. doi:10.48550/arXiv.1711.05225. URL https://doi.org/10.48550/arXiv.1711.05225.

[133] P. Rajpurkar, J. Irvin, et al. Deep learning for chest radiograph diagnosis: A retrospective comparison of the chexnext algorithm to practicing radiologists. *PLOS Med*, 15(11), 2018. doi:10.1371/journal.pmed.1002686. URL https://doi.org/10.1371/journal.pmed.1002686.

[134] Y. Bengio, J. Louradour, et al. Curriculum learning. *Proceedings of the 26th Annual International Conference on Machine Learning*, 2009. doi:10.1145/1553374.1553380. URL https://doi.org/10.1145/1553374.1553380.

[135] J. Deng, W. Dong, et al. ImageNet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009. doi:10.1109/cvpr.2009.5206848. URL https://doi.org/10.1109/cvpr.2009.5206848.

[136] X. Wang, Y. Peng, et al. Chestx-Ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. doi:10.1109/cvpr.2017.369. URL https://doi.org/10.1109/cvpr.2017.369.

[137] RSNA pneumonia detection challenge. https://www.kaggle.com/c/rsna-pneumonia-detection-challenge/data, 2018.

[138] M.D. Li, N.T. Arun, et al. Automated assessment and tracking of COVID-19 pulmonary disease severity on chest radiographs using convolutional siamese neural networks. *Radiol Artif Intell*, 2, 2020. doi:10.1148/ryai.2020200079. URL https://doi.org/10.1148/ryai.2020200079.

[139] J. Irvin, P. Rajpurkar, et al. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *Proceedings of the AAAI Conference on*

Artificial Intelligence, 33(01):590–597, 2019. doi:10.1609/aaai.v33i01.3301590. URL https://doi.org/10.1609/aaai.v33i01.3301590.

[140] M.A. Warren, Z. Zhao, et al. Severity scoring of lung oedema on the chest radiograph is associated with clinical outcomes in ARDS. *Thorax*, 73, 2018. doi:10.1136/thoraxjnl-2017-211280. URL https://doi.org/10.1136/thoraxjnl-2017-211280.

[141] L. McInnes, J. Healy, and J. Melville. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint*, 2018. doi:10.48550/arXiv.1802.03426. URL https://doi.org/10.48550/arXiv.1802.03426.

[142] S. Holm. A simple sequentially rejective multiple test procedure. *Scand J Stat*, 6: 65–70, 1979. URL https://www.jstor.org/stable/4615733.

[143] Chicago Department of Public Health. SARS-CoV-2 Variants. https://www.chicago.gov/city/en/sites/covid-19/home/sars-cov-2-variants.html, 2023.

[144] Centers for Disease Control and Prevention. ICD-10-CM Official Coding and Reporting Guidelines April 1, 2020 through September 30, 2020. https://www.cdc.gov/nchs/data/icd/covid-19-guidelines-final.pdf, 2020.

[145] D. Driggs, I. Selby, et al. Machine learning for COVID-19 diagnosis and prognostication: Lessons for amplifying the signal while reducing the noise. *Radiol Artif Intell*, 3 (4):e210011, 2021. doi:10.1148/ryai.2021210011. URL https://doi.org/10.1148/ryai.2021210011.

[146] J. Goncalves, L. Yan, et al. Li yan et al. reply. *Nat Mach Intell*, 3:28–32, 2021. doi:10.1038/s42256-020-00251-5. URL https://doi.org/10.1038/s42256-020-00251-5.

[147] X. Yang, X. He, et al. Covid-CT-dataset: A CT scan dataset about COVID-19. *arXiv preprint*, 2020. doi:10.48550/arXiv.2003.13865. URL https://doi.org/10.48550/arXiv.2003.13865.

[148] A. Kurakin, I. Goodfellow, and S. Bengio. Adversarial examples in the physical world. *arXiv preprint*, 2016. doi:10.48550/arXiv.1607.02533. URL https://doi.org/10.48550/arXiv.1607.02533.

[149] P. Rajpurkar, A. Joshi, et al. Chexpedition: Investigating generalization challenges for translation of chest x-ray algorithms to the clinical setting. *arXiv preprint*, 2020. doi:10.48550/arXiv.2002.11379. URL https://doi.org/10.48550/arXiv.2002.11379.

[150] S. Minaee, Y. Boykov, et al. Image segmentation using deep learning: A survey. *IEEE Trans Pattern Anal Mach Intell*, 44(7):3523–3542, 2022. doi:10.1109/TPAMI.2021.3059968. URL https://doi.org/10.1109/TPAMI.2021.3059968.

[151] C.E. Cardenas, J. Yang, et al. Advances in auto-segmentation. *Seminars in Radiation Oncology*, 29(3):185–197, 2019. ISSN 1053-4296. doi:10.1016/j.semradonc.2019.02.001. URL https://doi.org/10.1016/j.semradonc.2019.02.001. Adaptive Radiotherapy and Automation.

[152] Y. Fu, Y. Lei, et al. A review of deep learning based methods for medical image multi-organ segmentation. *Phys Med*, 85:107–122, 2021. ISSN 1120-1797. doi:10.1016/j.ejmp.2021.05.003. URL https://doi.org/10.1016/j.ejmp.2021.05.003.

[153] X. Zhou, R. Takayama, et al. Deep learning of the sectional appearances of 3d ct images for anatomical structure segmentation based on an fcn voting method. *Med Phys*, 44(10):5221 – 5233, 2017. doi:10.1002/mp.12480. URL https://doi.org/10.1002/mp.12480.

[154] W.D. Heaven. Hundreds of AI tools have been built to catch covid. None of them helped. MIT Technology Review, 30 July 2021 https://www.technologyreview.com/2021/07/30/1030329/machine-learning-ai-failed-covid-hospital-diagnosis-pandemic/, 2021.

[155] L. Wynants, B. Van Calster, et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. *BMJ*, page m1328, 2020. doi:10.1136/bmj.m1328.

[156] MIDRC. MIDRC. Rapid Response to COVID-19 Pandemic. MIDRC, 8 November 2023 https://www.midrc.org/, 2023.

[157] S.G. Armato III, K. Drukker, and L. Hadjiiski. AI in medical imaging grand challenges: Translation from competition to research benefit and patient care. *Br J Radiol*, 96 (1150), 2023. doi:10.1259/bjr.20221152.

[158] A. Krizhevsky, I. Sutskever, and G.E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J. Burges, et al., editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

[159] MIDRC. COVIDx-Challenge. https://www.midrc.org/events/covidx-challenge, 2022.

[160] K. Choudhary, B. DeCost, et al. Recent advances and applications of deep learning methods in materials science. *npj Comput Mater*, 8(1), 2022. doi:10.1038/s41524-022-00734-6. URL https://doi.org/10.1038/s41524-022-00734-6.

[161] D. Wang, E. Shelhamer, et al. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=uXl3bZLkr3c.

[162] V.R. Joseph. Optimal ratio for data splitting. *Stat Anal Data Min*, 15(4):531–538, 2022. doi:10.1002/sam.11583. URL https://doi.org/10.1002/sam.11583.

[163] R.R. Picard and K.N. Berk. Data splitting. *Am Stat*, 44(2):140, 1990. doi:10.2307/2684155. URL https://doi.org/10.2307/2684155.

[164] G. Afendras and M. Markatou. Optimality of training/test size and resampling effectiveness in cross-validation. *J Stat Plan Inference*, 199:286–301, 2019. doi:10.1016/j.jspi.2018.07.005.

[165] Q.H. Nguyen, H.B. Ly, et al. Influence of data splitting on performance of machine learning models in prediction of shear strength of soil. *Math Probl Eng*, 2021:1–15, 2021. doi:10.1155/2021/4832864. URL https://doi.org/10.1155/2021/4832864.

[166] H. Whitney, K. Drukker, et al. Role of sureness in evaluating AI/CADx: Lesion-based repeatability of machine learning classification performance on breast MRI. *Med Phys*, 08 2023. doi:10.1002/mp.16673. URL https://doi.org/10.1002/mp.16673.

[167] R. M. Haralick, K. Shanmugam, and I. Dinstein. Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-3(6):610–621, November 1973. doi:10.1109/TSMC.1973.4309314. URL https://doi.org/10.1109/TSMC.1973.4309314.