THE UNIVERSITY OF CHICAGO

INVESTIGATION OF COMPUTER VISION AND DEEP LEARNING ON THORACIC CT FOR ASSESSMENT AND EVALUATION OF CORONARY ARTERY CALCIUM, EMPHYSEMA, AND COVID-19

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
JORDAN FUHRMAN

CHICAGO, ILLINOIS
DECEMBER 2022

To my family for their endless love and support.

TABLE OF CONTENTS

LIST OF FIGURES vi

LIST OF TABLES xi

ACKNOWLEDGMENTS xii

ABSTRACT xiv

1 INTRODUCTION 1
- 1.1 Computed Tomography 2
- 1.2 Computer-Aided Diagnosis 3
- 1.3 Artificial Intelligence and Medical Imaging 4
- 1.4 Convolutional Neural Networks 5
- 1.5 Transfer Learning 6
- 1.6 Challenges for AI in CT 8
- 1.7 Explainability and Interpretability of AI in Medical Imaging 9

2 LUNG SCREENING AND THE LOW-DOSE CT DATASET 14
- 2.1 The Rise of Lung Screening 14
- 2.2 Low-dose CT Acquisition and Dataset 16
- 2.3 Additional Disease Findings in LDCT Screening: Coronary Artery Calcium . 17
- 2.4 Additional Disease Findings in LDCT Screening: Emphysema 20

3 AUTOMATIC SEGMENTATION AND CLASSIFICATION OF CORONARY ARTERY CALCIFICATIONS ON LOW-DOSE THORACIC CT 23
- 3.1 Coronary artery calcium 23
- 3.2 Revised U-Net Architecture 24
- 3.3 Training, Testing, and Statistical Analyses 25
- 3.4 ROC Analysis and Severity Evaluation 27
- 3.5 Evaluation of Severe Misclassification 28
- 3.6 Example Segmentations 31
- 3.7 Discussion of CACU-Net and Future Directions 31

4 EMPHYSEMA CHARACTERIZATION THROUGH MULTIPLE INSTANCE TRANSFER LEARNING IN LUNG SCREENING CT SCANS 35
- 4.1 Multiple Instance Learning 35
- 4.2 Transfer Multiple Instance Learning 36
- 4.3 Training, Testing, and Statistical Analyses 38
- 4.4 Binary Classification Performance 40
- 4.5 Attention Weight Interpretability Analysis 40
- 4.6 Discussion and Future Directions 43

5 PROGNOSIS AND TREATMENT RECOMMENDATION FOR COVID-19 PATIENTS THROUGH DEEP LEARNING ON THORACIC CT 47
5.1 COVID-19 and Medical Imaging 47
5.1.1 COVID-19 Databases 49
5.2 Preliminary Study: Cascaded Transfer Learning with COVIDSet1 52
5.2.1 Results from ROC Analysis and Temporal Analysis 57
5.3 Further Study: Deep Learning with COVIDSet2 62
5.3.1 Transformer Architectures 64
5.3.2 Incorporation of Human-Engineered Features and Clinical Presentation of COVID-19 64
5.3.3 Statistical Data Analyses 70
5.4 Results on COVIDSet2 70
5.5 Discussion and Conclusions 76

6 SUMMARY AND FUTURE DIRECTIONS 81

REFERENCES 85

LIST OF PUBLICATIONS AND PRESENTATIONS 100

LIST OF FIGURES

1.1 CT scan acquisition. (a) view within the CT scanner bore with x-rays traveling from the source, through the patient, and detected by the detector ring. Multiple acquisitions must be acquired by rotation of the x-ray source. (b) Scanning of different heights of the body by sliding the patient bed through the detector ring during acquisition. 2

1.2 Examples of CT sections for a variety of anatomical regions. Featuring the (a) head, (b) upper thorax, (c-d) mid thorax with images presented with different window levels to emphasize/diminish certain information, (e) abdomen. 4

1.3 Depictions of changing parameters in (a) convolutional filter size, (b) stride, and (c) padding. Selections in each of these parameters can heavily impact model performance and generalizability. For example, larger convolutional filters can capture a larger area (receptive field) of the input image but require increased parameterization that exposes the model to overfitting. Other selections can impact network structure and may restrict or enhance the type of task for which an architecture can be applied. 7

1.4 Example activation functions common to neural networks, including (left) linear activation, (middle) rectified linear unit (ReLU) activation, and (right) hyperbolic tangent activation. The selection of activation plays a role in the complexity of model which the network can achieve. As the layer output is determined from the activation function, activation selection may be dictated by prediction task, e.g., classification may require hyperbolic tangent or softmax activation while hidden layers may utilize ReLU activation. 8

1.5 Example questions regarding "explainability" and "interpretability" as used in this dissertation work. While the two are extremely similar, the intended audience and implementation of the model output is not equivalent [1]. 10

1.6 Portrayal of the tradeoff between learning performance, which is often associated with the number of learned parameters, and explainability. Note that deep networks are among the most common techniques for ML-based medical image evaluation, but also have generally low interpretability. There has been a strong push in recent years to develop techniques for general explanation of neural network predictions. Image acquired from Gunning (publicly available presentation with open distribution)[2]. 11

1.7 Examples of post-hoc heatmaps generated with Grad-CAM for a COVID-19 classification task. Note that while the heatmaps are sometimes readily interpretable as in (a), other examples are more difficult to understand, particularly when information outside the relevant anatomy or even outside the body area are determined to be influential. 13

2.1 Examples of CAC on LDCT scans. Green arrows indicate positive detections of CAC in the LMA (a) and in the RCA, CFX, and LAD in (b) while red arrows indicate false detections in the aortic root (a), aortic valve (b), and ribs (a-b) and in the descending aorta in (a). 19

2.2 Examples of emphysema phenotypes on CT scans. (a) shows CLE, with patchy, hypoattenuated regions throughout the upper lungs, (b) shows PSE at the periphery of the lung with discrete hypoattenuation, and (c) shows PLE, with widespread, homogeneous hypoattenuation covering a large portion of the lower lungs [3] 21
2.3 Proposed pipeline following LDCT acquisition. Note that the automatic evaluation is tailored to the specific condition based on desired output. 22

3.1 Standard U-Net architecture demonstrating an example segmentation of CAC within a CT section. The U-Net is composed of an encoding path, which utilizes convolutional and pooling layers to form an encoded representation of the input 2D image. The second half of the network then produces a binary segmentation map by decoding the representation into a probability map indicating if each pixel belongs to the desired structure/anatomy. 24
3.2 Schematic of the proposed deep network architecture, CACU-Net. Note that the softmax classification layers labelled with loss functions L1 and L2 attempt to match target outputs, with L1 aiming to segment all lesions and L2 aiming to segment lesions in each artery branch with different channels represents the different artery branches. A threshold was then applied to the L1 output to provide a binary segmentation which was then multiplied in an elementwise manner with each of the artery class probability maps produced by the L2 prediction layer. Artery class was then determined on a per-pixel basis by the maximum artery class probability for each pixel (LMA, RCA, CFX, LAD, and None). 26
3.3 Confusion matrices for the individual coronary arteries comparing predicted and reference standard ordinal scores for each of the four main coronary artery branches. 29
3.4 Confusion matrices for the scan level MSOS comparison with clinically relevant partitions of no (score: 0), mild (score: 1-3), and marked (score: 4-12) severity. Note that this is the confusion matrix of only one of the five models, but is representative of the results from all five. 30
3.5 Cropped example images and their corresponding automatically produced segmentations. Colors indicate artery classification, with green=LAD, red=LMA, blue=RCA, cyan=CFX. Image pairs (a-d) display examples of successful identification and classification for each artery while image pairs (e-h) demonstrate common cases of partial or complete failure in either segmentation (e-f) or classification (g-h). The red arrow in (f) marks an RCA lesion that was completely missed. 32

4.1 Model workflow of the Transfer AMIL approach. This includes feature extraction of CT images through an ImageNet pre-trained model based on methods developed by Antropova et. al. followed by attention-based MIL pooling based on methods developed by Ilse et. al. [4, 5]. Two outputs are generated for each LDCT scan input, the attention weights which identify influential slices for the classification task and the scan prediction for the presence of emphysema. 39

4.2 Attention weight curves illustrating the fit of attention weights from CT slices as a function of height in the lungs for (top) positive (red) and negative (green) LDCT scans and (bottom) for different dominant phenotypes of emphysema: centrilobular (blue), panlobular (pink), and paraseptal (turquoise). Since patients' CT scans have variable number of slices covering the lung region, in these plots, the range has been normalized to fit between Lung Top and Lung Bottom. . . . 42
4.3 Evaluation of common thoracic imaging features. The prevalence of each feature within the entire CT scan as identified by a radiologist and when selected by the top-k attention weighted slices. Note key differences between whole slice prevalence and selected prevalence: bronchial disease and architectural distortions were more heavily weighted while ground glass opacities are diminished. Further, the consistent representation across the top-k slices for different k demonstrates the model's tendency to more heavily weight slices with similar extracted representations. 44

5.1 Kaplan-Meier survival analysis assessing the duration of hospitalization with changing treatment and initial PSI score. In general, patients who received steroid treatments were hospitalized for longer periods of time, with particularly long stays for patients with more severe initial symptoms. This is expected, as more severe cases require increased treatment and recovery time. 51
5.2 (a) Schematic of the pretrained VGG19 network feature extraction approach operating on a 2D CT section. Max pooling layer features with the given dimensions were averaged and concatenated to produce a representative feature vector for each slice. (b) Full cascaded transfer learning workflow for pre-treatment assessment and during-treatment monitoring analysis. The feature extraction scheme displayed in (a) is utilized at the "Deep Transfer Learning: VGG19 Feature Extraction" stage of (b). 55
5.3 (a) The ROC curve demonstrating the classification ability of the cascade transfer learning method for estimating the likelihood that a COVID-19 patient would be recommended for steroid treatment or not. AUC $= 0.85 + / - 0.10$ with the accompanying 95% TPF confidence interval. (b) Distribution of deep learning scores of those patients who received steroids and those who did not. Note, this was obtained only based on the initial CT scan. Based on this plot, the method suggests steroid administration more frequently than the experienced intensivist (using a cutoff of 0.5). The red lines denote the median scores, the blue boxes include 50% of scores, while the black whiskers include all scores within 2σ of the mean. (c) Further demonstration of the separation/overlap of the deep learning score between the two classes. 58

5.4 The SVM-output prediction score assessed temporally through least squares fits. The x-axis indicates the full duration of hospitalization, with Ti referring to the time of initial CT acquisition and Tf referring to the time of final CT acquisition, which generally occurred shortly before discharge. The shaded regions denote one standard deviation above and below the fit line. Intuitively, this figure follows the example training case discussed in Section 2.3 which had early timepoints after which steroids were utilized and late acquisitions after which no steroids were administered. 59
5.5 Depiction of the self-attention module utilized in transformers a) Demonstrates the generation of attention weights for slice 1 via mapping each input representation w to Q and K representations, then interacting the relevant Q representation (in this case, q1) with the K representations of other slice representations. b) Shows the aggregation of attention weights with each V representation to form the output representation of slice 1, c1. 65
5.6 Depiction of how transformers may be utilized for CT scan evaluation. By first embedding the images to a feature representation via a convolutional neural network, self-attention modules can be utilized in parallel and in series to evaluate the image features and form a classification decision. 66
5.7 Workflow for the extraction of SBR features from CT scans. We first segment the lung and COVID-19 infection within individual CT slices using a novel dual-headed U-Net model, then extract intuitive features including intensity-based and volume ratio features. The full list of features is provided in Table 5.3. 67
5.8 Full feature fusion workflow of the MIL, SBR, and TC feature pipeline for each of the 5 cross validation folds. The artificial neural network classifier can be replaced with individual classifiers for each feature type to visualize the prediction fusion approach. 71
5.9 Comparison of predictions in the task of predicting COVID-19 steroid administration based upon initial patient CT scan and Bland-Altman plots from different models in ablation studies with patients who received steroids (red) and those who did not (green). (a) Compares predictions between the two ResNet50 and DenseNet121 models trained with ImageNet pre-training and attention-based MIL pooling (Table 5.4). Note that in one of the 5 cross-validation folds, the DenseNet121 model failed to converge to a useful set of parameters and the model produced a constant output regardless of input. This is observed with the unusual line of points through the center of the plot. (b) Compares the MIL model with SBR model. While both achieved similar AUCs, the SBR model seems to be recall cases at a much lower rate. (c) Comparing the same MIL model with the TC model. The poor performance of the TC features is prevalent here with relatively little structure visualized here. (d) Comparing the MIL model to the same model with feature fusion. The fused model tends to slightly underpredict cases compared to the model using MIL features alone; this may potentially be credited to the inclusion of the SBR features in the fused model. 74

5.10 Comparison of predictions in the task of predicting COVID-19 steroid administration based upon initial patient CT scan and Bland-Altman plots from different models in ablation studies with patients who received steroids (red) and those who did not (green). (a) Compares predictions between the two ResNet50 and DenseNet121 models trained with ImageNet pre-training and attention-based MIL pooling (Table 5.4). Note that in one of the 5 cross-validation folds, the DenseNet121 model failed to converge to a useful set of parameters and the model produced a constant output regardless of input. This is observed with the unusual line of points through the center of the plot. (b) Compares the MIL model with SBR model. While both achieved similar AUCs, the SBR model seems to be recall cases at a much lower rate. (c) Comparing the same MIL model with the TC model. The poor performance of the TC features is prevalent here with relatively little structure visualized here. (d) Comparing the MIL model to the same model with feature fusion. The fused model tends to slightly underpredict cases compared to the model using MIL features alone; this may potentially be credited to the inclusion of the SBR features in the fused model. 75
5.11 Example of two cases with successful prediction. (top) Patient received steroids. Investigating features reveals potentially contributing factors, including age, relatively small difference between mean disease and mean lung value, and large relative volume of disease. (bottom) Patient did not receive steroids, and based on the feature characteristics of extremely little diseased tissue volume and large difference between diseased and total lung pixel values, it becomes more clear why the model successfully reached the negative prediction decision. 79

LIST OF TABLES

1.1 Common Hounsfield values 3

2.1 LDCT database information 16
2.2 MSOS and Agatston CAC score comparison 19

3.1 Scan and Artery Performance AUCs (standard deviation). Bold indicates statistical significance. 28

4.1 Emphysema classification assessment 40
4.2 Quantitative attention weights from Figure 4.2 41

5.1 COVIDSet1 Database Information 50
5.2 COVIDSet2 Database Information 53
5.3 List of Incorporated SBR and TC Features for COVID-19 Steroid Prediction Task 68
5.3 Continued 69
5.3 Continued 70
5.4 Comparing ImageNet and RadImageNet Feature Extraction 72
5.5 Statistical testing for ImageNet vs. RadImageNet (p-values); bold indicates statistical significance 72
5.6 Transformer pooling vs. attention-based pooling 72
5.7 Incorporating additional feature types for steroid treatment classification prediction 73
5.8 Statistical testing for incorporation of additional feature; bold indicates statistical significance; Model numbers listed in Table 5.6 73

ACKNOWLEDGMENTS

So many individuals have contributed to my personal and professional growth, ultimately resulting in the development of this dissertation work. While I am thankful for the many who played a part in making this experience so enjoyable, there are several individuals without who this work would not have been possible.

Foremost, my advisor Maryellen Giger has been an incredible mentor and absolute pleasure to work with. I can not thank her enough for her professional guidance, kindness, patience, and support throughout the completion of this dissertation work. She and my other thesis committee members, Sam Armato, Patrick La Riviere, Lydia Chelala, and Heber MacMahon, provided valuable feedback several times over the past 5 years, thus I would like to thank them for their contributions.

It has also been a pleasure to interact with the many Giger lab members at lab meetings, conferences, and other events throughout the years. I would like to thank them for their suggestions, guidance, and friendship. To Hui Li, Karen Drukker, Chun-Wai Chan, Feng Li, Li Lan, Sasha Edwards, John Papaioannou, and Heather Whitney, thank you. I've also had the pleasure of mentoring many undergraduate researchers during my time at UChicago; thanks to Elise and Beatrice Katsnelson, Marlin Keller, Dallas Tada, Fernando Augustin-Elestario, Caitlin Huettl, and Peter Halloran.

All of the current and many past members of the Graduate Program in Medical Physics have supported me in a variety of ways during my time here. From professional collaboration and discussion to personal friendships and experiences, so many have positively impacted my UChicago experience. While my interaction with many was limited due to the COVID-19 pandemic, I must particularly thank the members of my cohort, Brittany Broder, Inna Gertsenshteyn, and Isabelle Hu, as well as the many other GPMP members who were excellent role models in the office and friends out of the office: Scott Trinkle, Adam Hasse, Sam Hendley, Talon Chandler, Kayla Robinson, Joe Foy, Corey Smith, Lindsay Douglas, Ben

Preusser, Natalie Baughn, Hadley DeBrosse, Linnea Kremer, Mena Shenouda, Mira Liu, Julian Bertini, Joseph Cozzi, and the many other past and present GPMP students who I wish I could have interacted with more.

My two Cool Office mates played a particularly important role in my doctoral experience. To Jennie Crosby, thanks for helping get everything started, initiating so many duck pond and $1 milkshake trips, and so many great memes. To Madeleine Durkee, thanks for more coffee and science/life discussions than I can count, for being a running/workout accountability buddy for many months, and for making me realize how great cats are.

Outside the GPMP, I've been able to make so many friends (including many who have already been listed!) who I can't thank enough for the many game nights, football and TV show watch parties, soccer games, nights out, nights in, and everything else we did along the way. Thanks Katie Dixon, Alex Nepon, Craig DeValk, Michelle Yelaska, Mahmoud Abouelnage, Chad Heer, Dalton and Mattie Moore, Mel Yamsek, Cassidy McPherson, Andrew Rohm, and many others.

Last but not least, I would like to thank my family for their unwavering support and encouragement. I would never have been able to reach this goal without them and I can not thank them enough for everything they do for me. Thanks Tim, Dee, Jackson, and Jared for everything (and of course, my cat Trudy for getting me through the pandemic!).

ABSTRACT

Over the past several years, new advances in computing hardware and artificial intelligence techniques have allowed deep learning to rapidly develop as a key tool in a broad range of fields. In medical imaging, significant attention has been devoted to exploring how these technologies can improve radiological workflow, including more efficient and more accurate image reading and serving as a rapid, objective reader acting concurrently with human radiologists. However, several challenges exist in applying typical deep learning technologies to CT scans. In this dissertation research, we consider three thoracic CT use cases and evaluate novel deep learning techniques to improve clinical utility.

The first aim of this dissertation was to develop a deep learning algorithm to evaluate coronary artery calcification (CAC) on low-dose thoracic CT (LDCT) scans. Coronary heart disease is the leading cause of death globally and CAC scores serve as a strong predictor for adverse events related to coronary heart disease. To automatically score LDCT scans, we developed a novel image segmentation network, CACU-Net, which identifies CAC on LDCT scans and classifies lesions based on the coronary artery branch. CACU-Net was able to identify which LDCT scans and individual arteries contain CAC and classify scans into clinically relevant categories based on severity of CAC, outperforming similar segmentation approaches.

A second algorithm was developed to detect emphysema on LDCT scans using a transfer attention-based multiple instance learning (TAMIL) approach. This novel technique evaluates slices individually using a transfer learning feature extraction algorithm that requires no additional network training. The slice features are then aggregated through a learned attention-based pooling method that both improves performance and provides interpretable information which a radiologist can utilize to understand model decision-making and identify cases in which the model may fail to perform. The TAMIL and CACU-Net pipelines have the

potential to be added to the screening clinical workflow for a rapid, objective augmentation of radiologist findings.

When the COVID-19 pandemic began in 2019 and CT served as a potential method of evaluation for severe COVID-19 patients, the techniques developed here were adjusted for COVID-19 evaluation. Thus, the final aim of this dissertation was to develop a multi-modal model which could aid clinicians in identifying when patients should undergo corticosteroid administration during their course of treatment. This algorithm included 1) a novel segmentation architecture, 2) an investigation of an improved TAMIL algorithm, and 3) comorbidity data. The proposed model demonstrates comparable classification performance compared to the unimodal variants with added interpretability. This technique could improve patient care during future waves of COVID-19, particularly in those patients that are immunocompromised and may require more aggressive treatment.

The research provided in these three aims has the potential to improve thoracic CT evaluation by providing more flexible, modality-appropriate models that may augment human readers at various stages of the clinical workflow. The application of such deep learning algorithms has significant potential to enhance clinical efficiency and to ultimately improve patient outcomes.

CHAPTER 1

INTRODUCTION

Over the past decade, new developments in computing hardware and artificial intelligence (AI) techniques have revolutionized the use of computers for a wide variety of medical tasks. While many groups had already developed computer-aided diagnosis (CAD) algorithms, the AI revolution garnered extreme interest in exploring applications in medical imaging ranging from relatively straightforward tasks, e.g., detecting visual presence of disease, to much more complex algorithms that can perform tasks with performance unmatched by any human reader, such as aggregating multispectral imaging data to identify patients at high risk for cancer. There are many factors that impact the development of AI techniques for medical imaging, including the amount and type of imaging data available, the clinical task in question, and the computing resources available. In this dissertation, we explore multiple thoracic conditions that may be evaluated with the use of AI. We first provide necessary background on machine learning technologies then briefly discuss lung cancer screening in Chapter 2 to motivate Chapters 3 and 4. Each of the three primary projects is then discussed:

- automatic segmentation and classification of coronary artery calcifications on low-dose thoracic CT (Chapter 3)
- emphysema characterization through multiple instance transfer learning in lung cancer screening CT (Chapter 4)
- prognosis and treatment recommendations for COVID-19 patients through deep learning on thoracic CT (Chapter 5)

In these aims, we attempt to develop robust, applicable methods that can be applied ultimately in radiology practices. We also investigate other key issues of CAD systems in radiology, such as model interpretability, and provide suggestions for future direction on

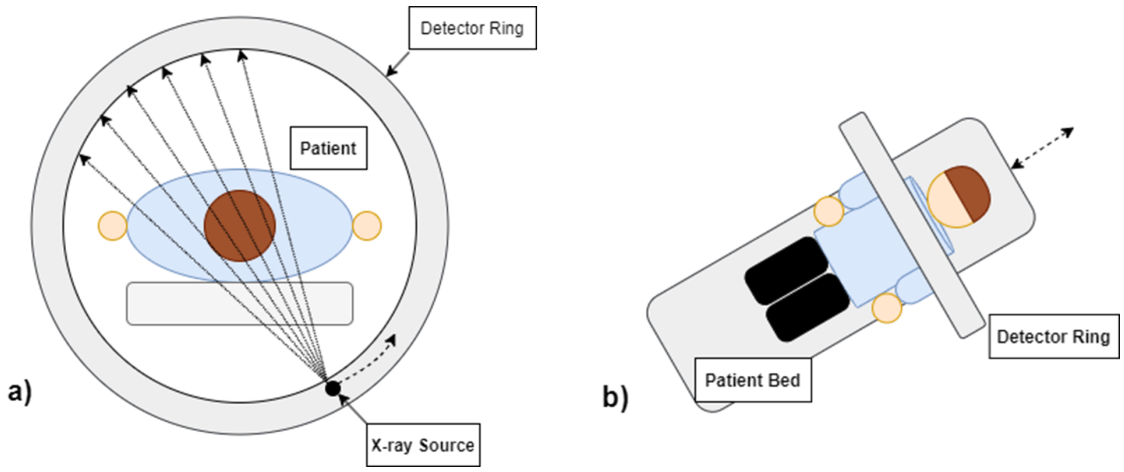

Figure 1.1: CT scan acquisition. (a) view within the CT scanner bore with x-rays traveling from the source, through the patient, and detected by the detector ring. Multiple acquisitions must be acquired by rotation of the x-ray source. (b) Scanning of different heights of the body by sliding the patient bed through the detector ring during acquisition.

each of these topics. The work presented in this dissertation lies at the intersection of imaging science, medical physics, medicine, and computer science, thus, this introductory chapter provides background knowledge of AI techniques, with focus on supervised deep learning for computer vision applied for CT scan evaluation.

1.1 Computed Tomography

In 1972, computed tomography (CT) became the first medical imaging modality to provide slice-based compositional information of the internal human body in the clinic [6]. This technology revolutionized medical imaging and plays a critical role in modern medicine, including radiology [6–8]. Briefly, CT images are acquired by detecting x-rays after projection through the human body at several angles and reconstructing voxel information in the form of Hounsfield Units (HU), which provide compositional information in each voxel based on the linear attenuation coefficient of the material in that voxel (Figure 1.1) [6–9].

Following the HU scale, image readers are able to relate CT image voxels to human anatomy/pathology by comparison to standard HU values for different anatomies (Table 1.1) [9]. The foundational work for the development of CT imaging resulted in the joint awarding of the 1979 Nobel Prize for Medicine to Hounsfield and Cormack [9–11].

Table 1.1: Common Hounsfield values

Tissue/Material	*Hounsfield Unit Value*
Air	-1000
Lung	-600
Fat	-100
Water	0
Soft Tissue	20 to 40
Blood	40
Calcium/Bone	100+
Metal	3000+

In the 50 years since its inception, CT imaging technology has progressed immensely, providing images faster, at a higher quality, and with reduced radiation risk to the patient [12]. Because of this, CT is now a powerful, flexible tool for radiological evaluation and is a primary cog in the management of a variety of diseases and patient conditions (Figure 1.2).

1.2 Computer-Aided Diagnosis

In addition to improvements in CT image acquisition and reconstruction, new technologies in CAD have recently emerged that can aide radiologists in accurate, consistent image reading. While these technologies range from intuitive, human-engineered methods to complex, large-scale deep learning approaches, their key goal is to improve a reader's (e.g., a radiologist's) ability to provide objective, consistent diagnoses through quantitative image analysis. CAD systems have been implemented with increasing frequency over the last decade as computational hardware and algorithms have improved, a trend that is likely to continue in coming years. Many CAD technologies stem from non-medical applications, such as natural language processing (NLP), natural image classification, and bioinformatics [13–16]. How-

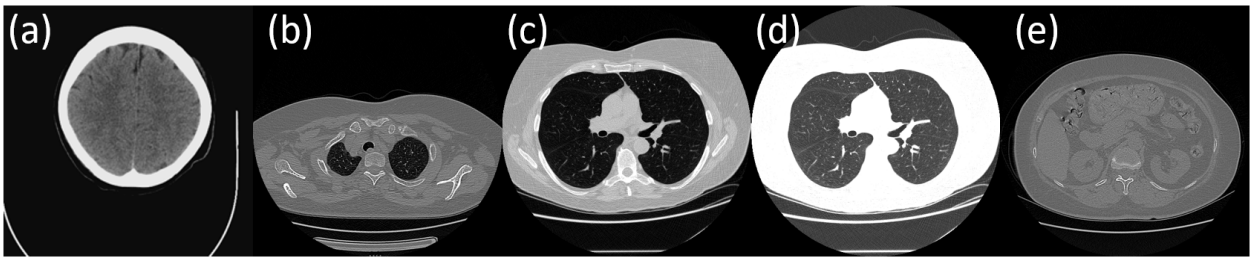

Figure 1.2: Examples of CT sections for a variety of anatomical regions. Featuring the (a) head, (b) upper thorax, (c-d) mid thorax with images presented with different window levels to emphasize/diminish certain information, (e) abdomen.

ever, there are several challenges in applying these technologies in medical image evaluation, including ethical considerations, data restrictions, and differences in image content [17–19]; for example, the physical context of pixel information and the 3-dimensional aspect of CT are often not compatible with standard natural image machine learning architecture. Thus, the primary goal of this dissertation is to explore novel CAD systems applied to CT imaging for a variety of disease evaluations with a focus on thoracic abnormalities.

1.3 Artificial Intelligence and Medical Imaging

AI broadly describes a variety of algorithms that attempt to mimic human intelligence and solve problems through pattern recognition [13], including those early methods used in CAD. In radiology practice, several tasks have been shown to be automatable through AI, including disease detection, prognostic prediction, and biomedical image segmentation, among others, stemming from recent significant advances made by the machine learning and medical imaging communities [1, 13, 20, 21]. These AI systems are trained through machine learning (ML) algorithms, which attempt to identify patterns within structured data, e.g., medical images. ML algorithms can be generally categorized as either supervised or unsupervised learning based on the method of model evaluation/adjustment during training. Supervised learning is much more common in medical imaging, following a paradigm in which each input datum has an associated reference label/class that the model aims to predict; during

training, model predictions are compared to the reference standard through a loss function and model parameters are updated based on the loss to improve performance for the given task. Supervised learning schemes are advantageous in many situations given that they are relatively straightforward to train and can produce high-performing, task-specific models. Alternatively, unsupervised learning algorithms do not focus on a particular prediction task; rather, they are developed by clustering the training data based on identified trends/similarities within the data without explicit task bias. Thus, unsupervised learning schemes tend to trade performance on a specific task for generalizability [22, 23]. Importantly, both paradigms require a large, representative training data set to achieve acceptable widespread implementation.

1.4 Convolutional Neural Networks

Currently, the most common type of ML model in medical image analysis is the convolutional neural network (CNN). This class of AI architectures utilizes learned convolutional filters as the backbone of the architecture; by utilizing several optimized convolutions both in parallel and in series in combination with other operations (e.g., pooling), a CNN is able to extract multi-scale quantitative features that can be used for predictive modelling. There are several CNN architectural design decisions that can play a role in performance and flexibility in data structure. For convolutional layers, these include filter size, stride, and padding, which are described with advantages and disadvantages in Fig 1.3, among others, while pooling layer selections impact the way in which feature maps are reduced, e.g., by either mean or argmax operations. Importantly, all layers contain an activation function, which injects nonlinearity into the model and plays a key role in allowable values during task prediction (Fig 1.4).

The number of parameters in typical medical imaging CNNs can range from $10^4 - 10^7$; this is generally far greater than the amount of data available, causing the model optimization problem to be underdetermined. As noted in Section 1.2, this causes CNNs to require large

amounts of well-annotated data to appropriately capture information relevant to the given task and avoid overfitting. To emphasize this, consider an AI model tasked with predicting presence/absence of lung cancer based on color information of a dermatological image (e.g., optical camera). Suppose that the patient cohort used for model training only contained patients with white/Caucasian skin color; regardless of architectural design selections and training paradigms (supervised vs. unsupervised learning), it is likely that the model would fail to generalize to non-white patients with strong performance. This type of bias caused by overfitting to the training data is incredibly common in machine learning, especially in the medical imaging domain in which data is relatively heterogeneous and data is limited.

1.5 Transfer Learning

While there are many strategies that attempt to minimize the likelihood of overfitting bias including regularization, class weighting, and feature selection, one proposed solution that has been widely implemented both in medical imaging and elsewhere is transfer learning [4, 13, 24–27]. Transfer learning describes the application of a model that was pre-trained for task T1 in domain D1 to a related task T2 in a potentially new domain D2. One assumes that the features learned for D1 will, to some extent, generalize to D2 and be applicable to the new task T2, providing improved baseline performance than a randomly parameterized model. Transfer learning is commonly utilized in medical imaging via pre-training for classification on the ImageNet database, a collection of millions of natural, everyday images [28, 29]. Strong performance on this database requires a deep learning model to characterize potentially generalizable information such as object shape, color intensity, etc., which are also likely to be important in the medical imaging task.

However, feature transfer is usually suboptimal compared to directly training in the task domain D2. Often, differences between the two domains can be problematic and may require image pre-processing that would otherwise be unnecessary or nonsensical for deep network

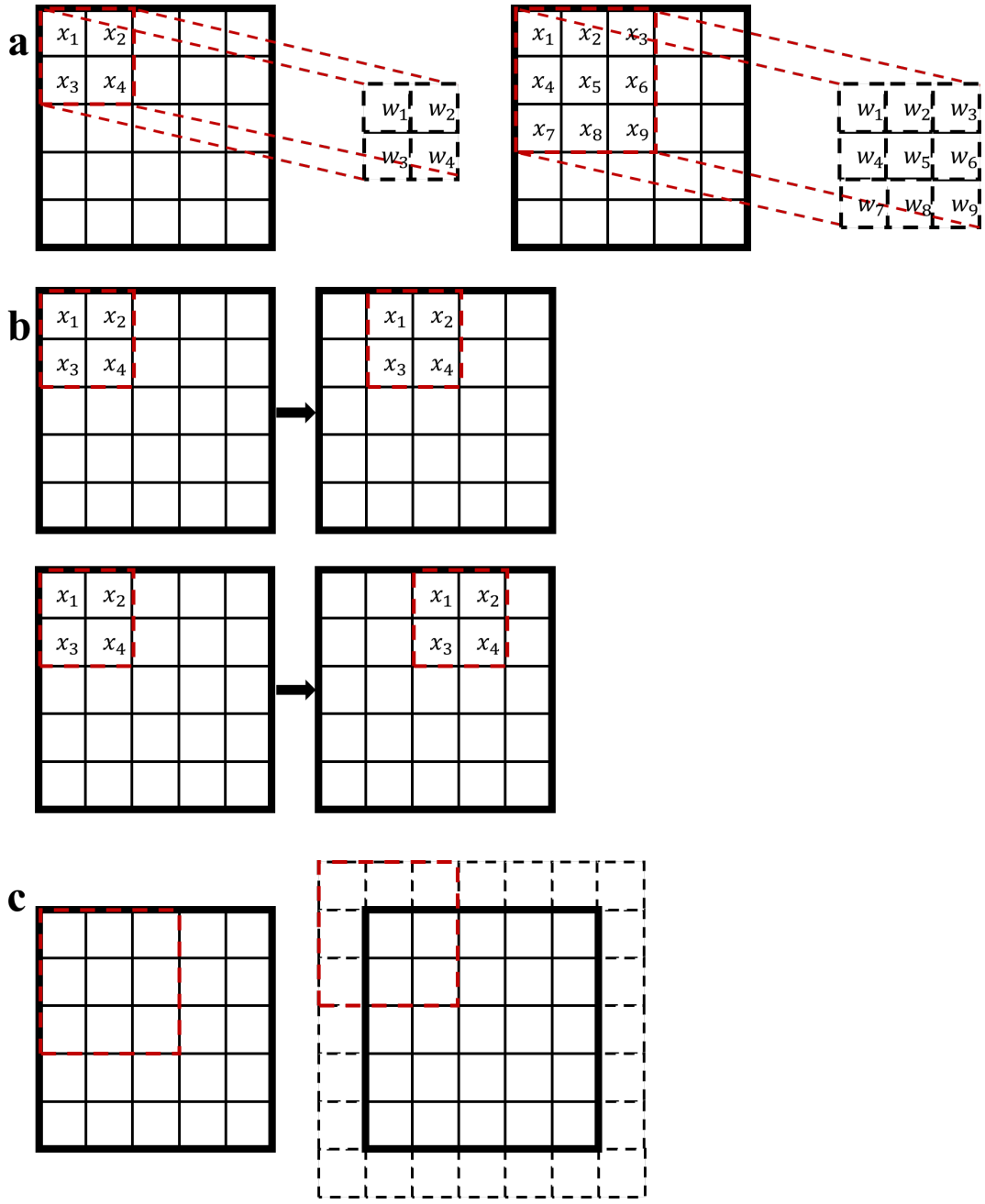

Figure 1.3: Depictions of changing parameters in (a) convolutional filter size, (b) stride, and (c) padding. Selections in each of these parameters can heavily impact model performance and generalizability. For example, larger convolutional filters can capture a larger area (receptive field) of the input image but require increased parameterization that exposes the model to overfitting. Other selections can impact network structure and may restrict or enhance the type of task for which an architecture can be applied.

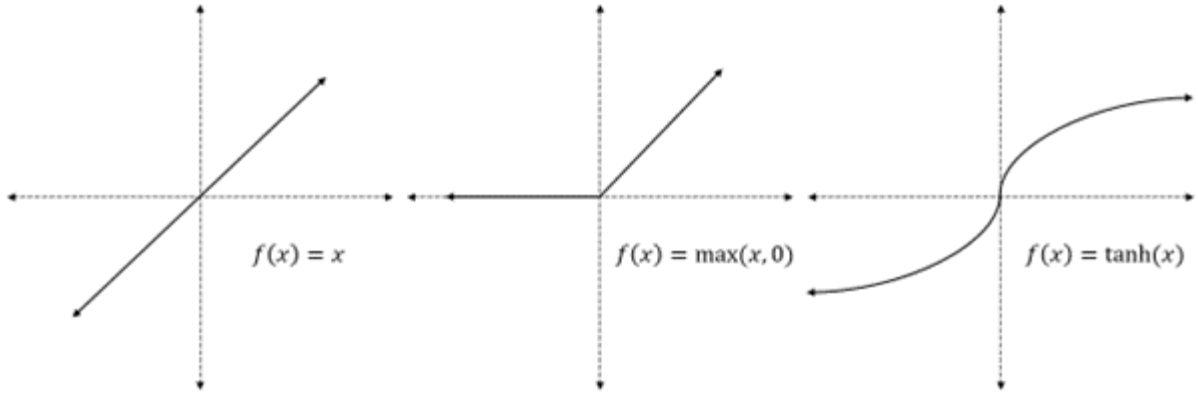

Figure 1.4: Example activation functions common to neural networks, including (left) linear activation, (middle) rectified linear unit (ReLU) activation, and (right) hyperbolic tangent activation. The selection of activation plays a role in the complexity of model which the network can achieve. As the layer output is determined from the activation function, activation selection may be dictated by prediction task, e.g., classification may require hyperbolic tangent or softmax activation while hidden layers may utilize ReLU activation.

analysis. Consider the prior example; the natural image domain D1 input space requires 3-channel (i.e., RGB) images as input but the D2 medical images are grayscale. The grayscale images must be converted to a 3-channel input (often by copying the grayscale image 3 times as different "channels"), but the 3-channel filters learned for the task T1 that consider the interaction between channel dimensions may be nonsensical for the task T2, leading to a suboptimal model.

In practice, there are several strategies that allow for improved transfer such as training a new classifier utilizing features extracted from the transfer learned model. This and other aspects of transfer learning will be discussed further in Chapters 4 and 5 during their specific application.

1.6 Challenges for AI in CT

Practically, there are several challenges that arise when machine learning techniques are applied to CT. The most obvious and potentially most impactful is that CT images are

3-dimensional (3D) while most other domains design models that are structured to only evaluate 2-dimensional (2D) information. While many investigators choose to simply apply 2-D models to individual CT sections, this fails to incorporate the rich 3-D information that may improve classification performance and generalization. Further, evaluating CT scans slice-by-slice is not clinically relevant unless the slice predictions can be sensibly aggregated to form a reasonable scan prediction. Additionally, even if models could be developed that incorporate fully 3D information, there are other issues that arise due to non-isotropic resolution and unequal image sizes between patients. These can cause problems during deep network construction and will be discussed in further detail in Chapters 4 and 5 where the proposed multiple instance model schemes may mitigate undesirable effects.

1.7 Explainability and Interpretability of AI in Medical Imaging

Another aspect of machine learning for medical imaging that has been of high interest in the past years is the lack of transparency in technology, including ML systems, contributing to critical decisions [30–32]. Because of the perception of ML algorithms as "black box" algorithms which require little or no explicit human intervention, it can be difficult to ethically justify their use in high-stakes decisions, especially because this type of technique lends little indication of when it is likely to fail [30–34]. Thus, the investigation of methods that can explain why an AI system provided a particular prediction is critically important.

In medical imaging, machine learning is typically applied to improve medical image assessment and workflow [13, 14, 21, 35–49]. The choice in ML method is dictated by the imaging task, which then influences which interpretability techniques may be appropriate. We reviewed several approaches that provide interpretable radiological AI systems.

The terms "explainability" and "interpretability" have been increasingly discussed in the AI community, particularly as they pertain to AI performance and ethics, and have raised several important questions [50–52]. Will radiologists more heavily weigh AI output with im-

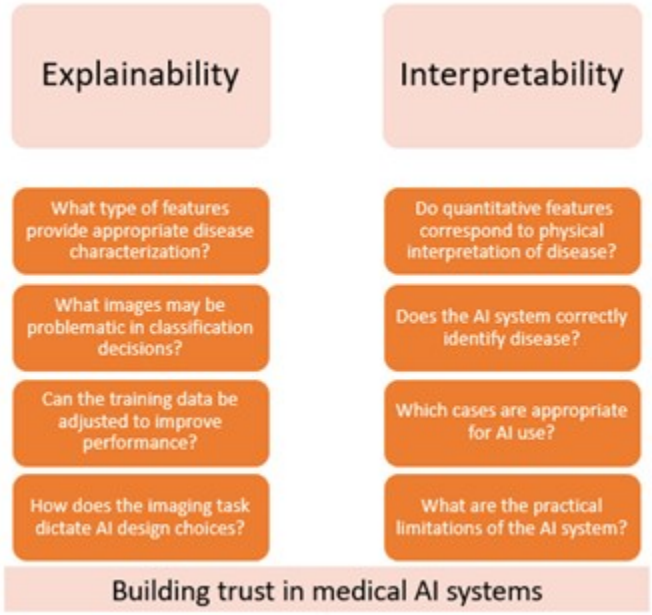

Figure 1.5: Example questions regarding "explainability" and "interpretability" as used in this dissertation work. While the two are extremely similar, the intended audience and implementation of the model output is not equivalent [1].

proved interpretability? Can the incorporation of explainable techniques also benefit model performance? Who is responsible when inappropriate decisions are made based on AI output? These and similar questions have instigated several attempts to define "explainability" and "interpretability" in AI; however, many definitions have considerable overlap or clash [50–52]. In this dissertation, we utilize "explainability" to refer to techniques applied by a developer or researcher to explain and improve the AI system while "interpretability" refers to understanding algorithm output for end-user implementation. Questions portraying the intended meaning of each term are given in Fig. 1.5 [1].

Several groups have provided extensive surveys of explainable AI and visualization [30, 51, 53–57]. However, these reviews focus on more general problems in both medical and non-medical disciplines (e.g., non-image assessments).

In general, a tradeoff exists between the complexity/depth of an AI system and its interpretability, with classical, shallow algorithms, such as decision trees, providing more ex-

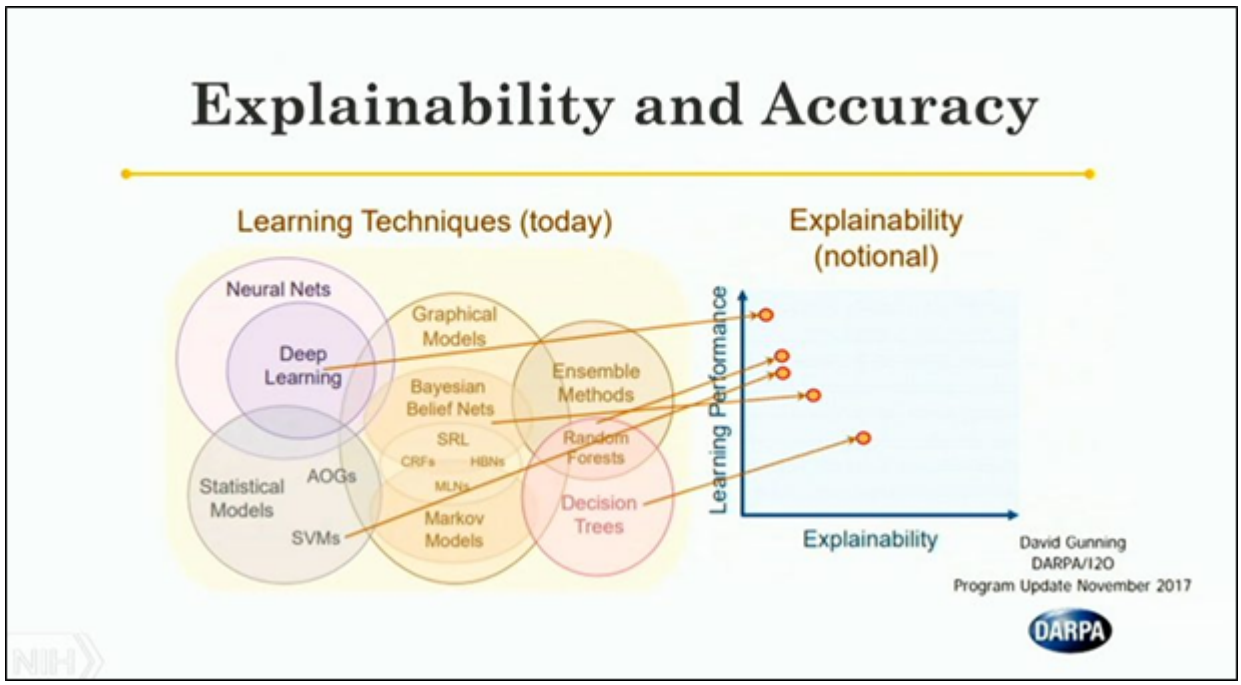

Figure 1.6: Portrayal of the tradeoff between learning performance, which is often associated with the number of learned parameters, and explainability. Note that deep networks are among the most common techniques for ML-based medical image evaluation, but also have generally low interpretability. There has been a strong push in recent years to develop techniques for general explanation of neural network predictions. Image acquired from Gunning (publicly available presentation with open distribution)[2].

plainable output with a potentially reduced performance [2, 57, 58]. Figure 1.6 depicts this phenomenon for several commonly used algorithms. It is important to note that finding the optimal operating point between system performance, which will improve patient management, and system interpretability, which will lead to more frequent implementation and trust in radiological practice, is critical.

AI interpretation techniques can be generally partitioned into two categories: post-hoc interpretability and inherent interpretability. During the initial rise of AI techniques in the 2010s, post-hoc techniques dominated the interpretability space due to limitations in computing and because of the success that several of these methods demonstrated, culminating with gradient-weighted class activation maps (Grad-CAM) in 2017 [59]. However, several studies have demonstrated failures of post-hoc methods, in particular showing non-sensical

explanations (e.g., identifying regions with no imaging content such as a zero-masked region as providing high influence to the classification decision) and have led many to distrust such methods (Fig 1.7).

More recently, interpretability methods that are inherent to model classification and performance have been of high interest to the AI community, specifically through attention modules. Attention modules are named as such because they are components of a machine learning architecture that attempt to guide the "focus" of a model to correctly identify signal that is relevant to the classification task. This is aptly demonstrated by the first successful use of attention applied to the U-Net architecture in which Oktay utilized attention to emphasize feature information relevant to the pancreas and diminish extraneous information. Inherent interpretability techniques have demonstrated such strong performance that entire networks can now be composed of these attention modules. In this dissertation, both post-hoc and inherent interpretability techniques are utilized in an attempt to understand model performance and identify problematic cases, which may be key in eventual clinical translation of such models. In this dissertation research, two different attention modules are utilized to improve model performance and provide interpretable output through attention-based multiple instance pooling (Section 4.5). The inclusion of such techniques allows for improved algorithm validation and trust and can be leveraged to increase the likelihood of clinical translation.

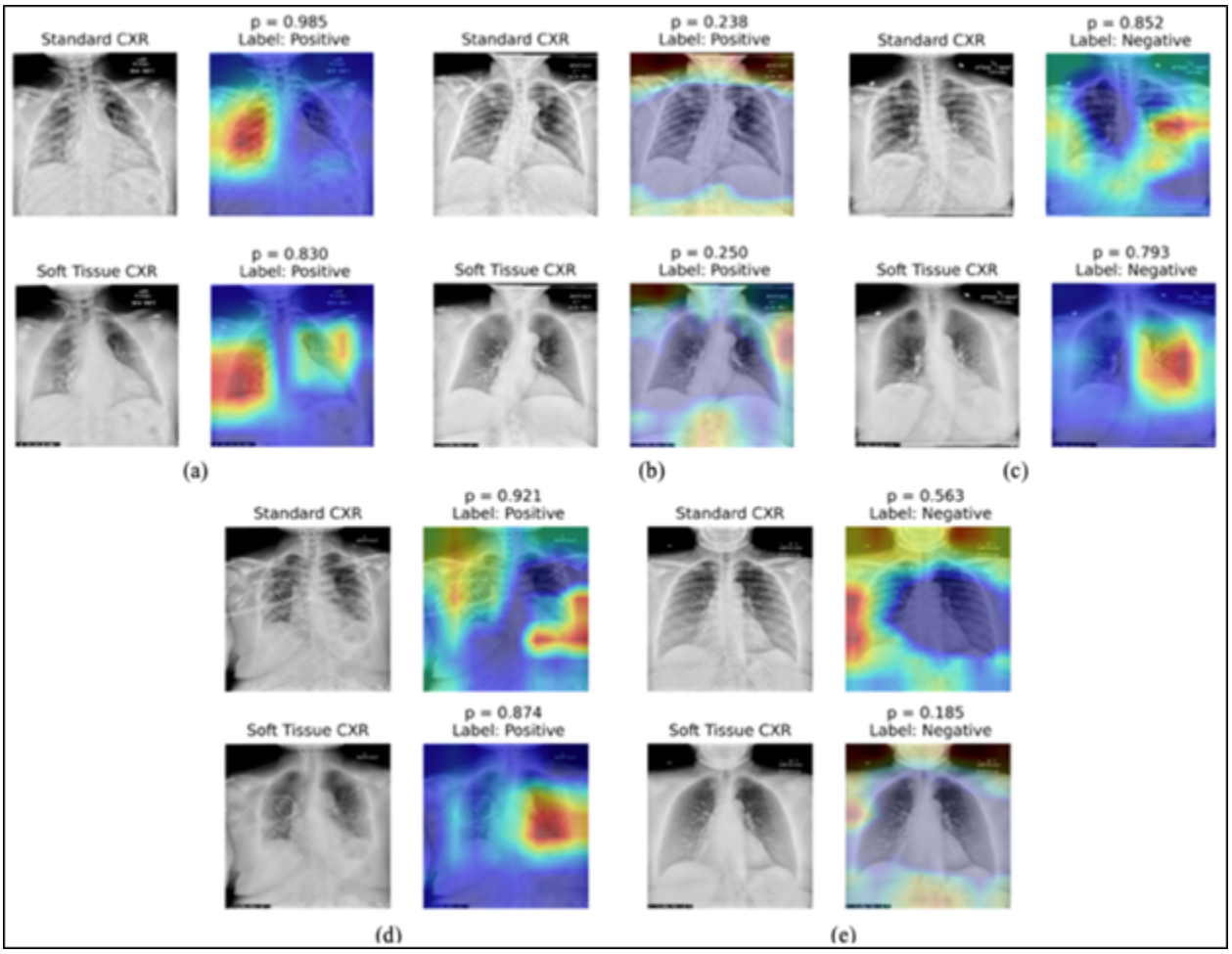

Figure 1.7: Examples of post-hoc heatmaps generated with Grad-CAM for a COVID-19 classification task. Note that while the heatmaps are sometimes readily interpretable as in (a), other examples are more difficult to understand, particularly when information outside the relevant anatomy or even outside the body area are determined to be influential.

CHAPTER 2

LUNG SCREENING AND THE LOW-DOSE CT DATASET

While the AI algorithms that have been developed for this study are generally applicable to any type of slice-based imaging modality, Chapters 3 and 4 of this dissertation utilize low-dose CT (LDCT) acquisitions that present unique challenges related to the given AI tasks. Because of this, it is important to understand the context of these images and why we choose to utilize LDCT scans for use cases.

2.1 The Rise of Lung Screening

The Center for Disease Control's (CDC) most recent report on cancer statistics reported that lung cancer was the leading cause of cancer death in 2019 at approximately 139,000 deaths with approximately 221,000 new cases [60]. One of the most important aspects of lung cancer mortality is the stage of detection, with earlier detection leading to a significantly improved chance of survival as the earlier stage generally increases the chances of curative treatment and positive outcomes [61–64]. The primary risk factor for lung cancer is significant smoking history, and while the number of heavy smokers has been reduced in recent years and led to reduced rates of lung cancer, there are still many current and former smokers that remain at an increased risk of lung cancer, particularly as they reach old age [65–67].

In the late 1990s, the Early Lung Cancer Action Program (ELCAP) evaluated annual lung cancer screening in high-risk populations via chest radiograph and LDCT acquisition in an attempt to identify non-calcified pulmonary nodules in patients with no prior cancer [63]. The results of this study found that the LDCT images significantly improved detection of such nodules at earlier stages, thus allowing for increased flexibility and probability of success in treating malignant findings [63]. This study initiated widespread interest in lung screening via LDCT with several other groups beginning large-scale investigations in the

mid-2000s, including the National Lung Screening Trial (NLST)[62], several European efforts (e.g., NELSON, LUSI, MILD)[68–70], and the expansion of the original ELCAP to include international members (I-ELCAP). In each of these follow-up studies, the screening trials demonstrated significantly reduced risk of lung cancer-associated mortality and, in the case of the NLST, reduced risk of overall mortality.

Recently, the United States Preventative Services Task Force (USPSTF) recommended expanded eligibility criteria for lung screening to include all individuals aged 50 to 80 years old (previously 55-80 years old) who have a smoking history of 20 pack-years (previously 30 pack-years) and are either current smokers or have quit within the past 15 years [71]. These individuals should be annually scanned by LDCT acquisition until they either no longer fit the criteria or another health problem significantly reduces the likelihood that lung cancer detection would benefit the patient. Considering this and the fact that national uptake of lung cancer screening remains low and there are continued efforts to increase implementation, it is highly likely that lung screening will experience widespread increases in implementation in the coming years.

While the primary goal of LDCT screening is to detect lung cancer, there are a variety of other thoracic diseases that can be visualized within the LDCT scan range, namely within the heart and lungs [72–75]. Radiologists evaluate for all potential diseases during reading, but this process is arduous and time-consuming. Given the rise of deep learning technology and applications in medical imaging, it is prudent to develop algorithms that can serve as concurrent readers and mitigate any reading errors or variability.

This chapter reviews the LDCT dataset utilized in Chapters 3 and 4 for the evaluation of additional findings in lung screening patients.

Table 2.1: LDCT database information

Number of Cases	865 (70%Train/10%Val/20%Test)
Dates of Acquisition	1997 - 2017
Sex at Birth	Male (384) Female (431) NA (50)
Smoking Status	Current (257) Former (469) Never (89) NA (50)
Age	Mean (66.8) SD (11.4) Range (33-39)
Pack-Years of Smoking	Mean (36.2) SD (30.8) Range (0-199)
Scanner Manufacturer	GE Medical Systems Siemens
Exposure Time	Range(250-2100)
kVp	(100, 120, 140)
Slice Thickness	Range (0.5mm - 10mm)
MSOS: None: 0	255 (31.3%)
Mild: 1-3	371 (45.6%)
Moderate: 4-6	114 (14.0%)
Severe: 7-12	74 (9.1%)
Emphysema: None:	500 (58.1%)
Mild/Moderate:	243 (28.3%)
Severe:	117 (13.6%)

2.2 Low-dose CT Acquisition and Dataset

While LDCT is utilized globally for lung screening, there is no established specific definition for what constitutes "low-dose" [76]. Most screening protocols suggest scan parameters between 120-140 kVp and 30-100 mAs; the I-ELCAP protocol, under which the data used in this research were acquired, is within this range at 120 kVp and 40 mAs at most [76]. All images utilized in this dissertation research were acquired with no contrast enhancement, no electrocardiogram-gating (ECG-gating), and reconstructed utilizing a standard kernel. Other information related to the acquisition of the LDCT scans is located in Table 2.1.

2.3 Additional Disease Findings in LDCT Screening: Coronary Artery Calcium

Coronary artery calcium (CAC) is typically scored following the Agatston criteria developed in 1990 [77]. This technique evaluates calcified plaques within the coronary arteries on CT slices of 3 mm thickness defined as lesions with peak value greater than 130 HU with an area greater than at least 1 mm^2. The Agatston score is then calculated by a weighted sum of lesion areas; the weighting factor is based on the peak density of a given lesion with the following ranges:

$$130 - 199\ HU = 1, \quad 200 - 299\ HU = 2, \quad 300 - 399\ HU = 3, \quad > 400\ HU = 4$$

Based on this weighted sum, the Agatston score has five clinically relevant categories:

$$None : 0, \quad Minimal : 1 - 10, \quad Mild : 11 - 100, \quad Moderate : 101 - 400, \quad Severe :> 400$$

Recent research has investigated more complex cardiac risk models, incorporating either the Agatston score or other CAC scoring techniques [78–80].

Currently there are very few options to reduce the risk of CAC, all of which have improved success when applied prior to severe progression [81, 82]. The most effective solution is lifestyle change during early stages that reduces the development of CAC, but other treatments include statins, anticoagulants, and more invasive procedures such as coronary angioplasty and coronary artery bypass graft (CABG). Unfortunately, CAC does not typically coincide with relevant symptoms until it reaches progressive disease stages, thus dedicated screening programs and additional risk factor disease detection are the primary methods of CAC scoring.

Additionally, coronary artery disease has several shared risk factors with lung cancer, including the two primary lung screening criteria of smoking history and old age [72, 80, 83–85]. It is thus fitting and appropriate to score CAC on LDCT scans for lung screening. However, the Agatston score may be inappropriate for application on LDCT scans because it was defined on only 3 mm scans acquired at standard dose with ECG-gating; the LDCT scans were acquired at 0.5 mm slice thickness with reduced dose and no cardiac gating [74].

Instead, an alternative ordinal scoring system, the Mount Sinai Ordinal Score (MSOS), can be utilized that has demonstrated comparable performance to the Agatston score while maintaining flexibility in application to LDCT scans [74, 86]. The MSOS is identified as an addition of severity scores assigned to each of the four main coronary artery branches: left main (LM), left anterior descending (LAD), circumflex (CFx), and right coronary artery (RCA). Examples of calcium present within each branch as well as other noted potential sources of calcium are presented in Fig 2.1. Each branch is assigned a score from 0-3 based on the extent of CAC within that branch, with a score of 0 corresponding to no CAC present, 1 corresponding to CAC in less than 1/3 of the length of the artery, 2 corresponding to between 1/3 and 2/3 of the artery length filled with CAC, and 3 corresponding to greater than 2/3 of the length presenting calcification. The sum of the branch scores provides the MSOS CAC score, ranging from 0-12. Similar to the Agatston score, the MSOS can be stratified into clinically relevant categories regarding the risk and treatment of coronary heart disease (Table 2.2).

The research utilized in Chapter 3 attempts to automatically provide the MSOS CAC score on LDCT scans. The prevalence of the MSOS categories in the database used for this research can be found in Table 2.1. Note that 51 cases were excluded from CAC evaluation due to evidence of prior knowledge of CHD either through visual presence of stent or CABG, resulting in a total of 814 cases utilized for the CAC AI research.

Table 2.2: MSOS and Agatston CAC score comparison

MSOS Range	Agatston Score Range	Clinical Recommendation
0	0	Low probability of CHD, no treatment recommended
1-3	1-100	Mild to moderate probability of CHD, potential use of statins
4-12	>100	High probability of CHD, use of statins and additional medications, potential invasive procedure required

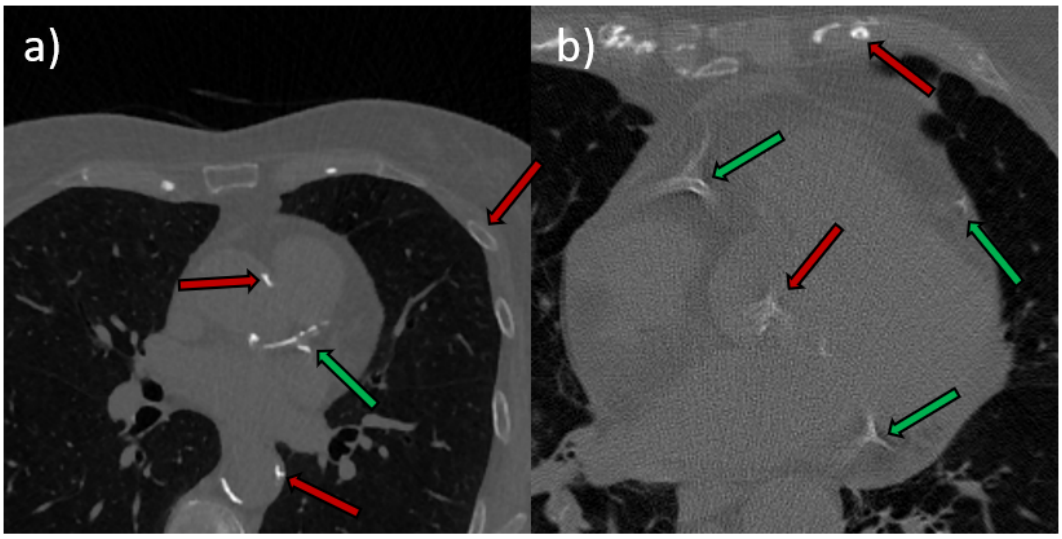

Figure 2.1: Examples of CAC on LDCT scans. Green arrows indicate positive detections of CAC in the LMA (a) and in the RCA, CFX, and LAD in (b) while red arrows indicate false detections in the aortic root (a), aortic valve (b), and ribs (a-b) and in the descending aorta in (a).

2.4 Additional Disease Findings in LDCT Screening: Emphysema

Emphysema is a lung condition in which the inner lining of the alveolar walls of the pulmonary lobule, the location of gas exchange within the lungs, break down, restricting the lungs' compliance and ability to contract properly [87]. Because of this, gas can no longer be expelled from the lungs, leaving no volume for oxygen to fill upon inhalation. The American Lung Association reports that over 2 million individuals have been diagnosed with emphysema, of which one of the primary causes is smoking history along with other genetic dispositions and air contaminants; thus, a significant proportion of the lung screening population is likely to have some degree of visible emphysema upon LDCT presentation [72, 85]. Similar to CAC, there is no current cure to emphysema, so early detection and treatment can significantly mitigate progression and improve patient quality of life.

On CT scans, airspace enlargement and the trapped air caused by a patient's inability to exhale properly appear as regions of hypoattenuation throughout the lungs, with prevalent regions of emphysema impacted by the dominant phenotype. The three phenotypes are centrilobular emphysema (CLE), panlobular emphysema (PLE), and paraseptal emphysema (PSE), examples of which are shown in Fig 2.2 [88–91]. In the case of a lung screening population, the most dominant phenotype is CLE, which is commonly found in asymptomatic elderly patients (similar to the motivation for lung cancer detection). CLE typically presents with an upper lung lobe predominance with a patchy distribution throughout the lungs and visual hypoattenuation in the central part of the pulmonary lobule. Alternatively, PLE has a lower lobe predominance with a more uniform distribution across the pulmonary lobules, and PSE predominates with consolidated hypoattenuation towards the periphery of the pulmonary lobules.

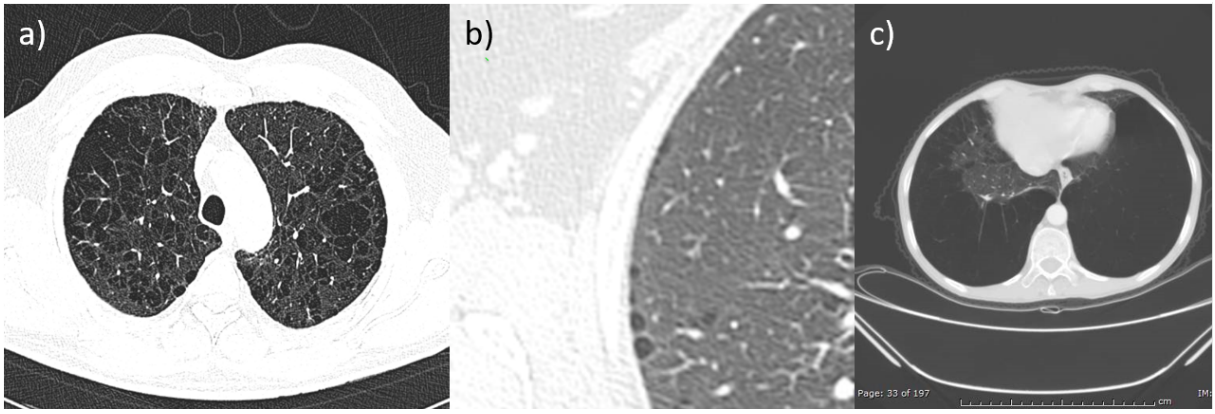

Figure 2.2: Examples of emphysema phenotypes on CT scans. (a) shows CLE, with patchy, hypoattenuated regions throughout the upper lungs, (b) shows PSE at the periphery of the lung with discrete hypoattenuation, and (c) shows PLE, with widespread, homogeneous hypoattenuation covering a large portion of the lower lungs [3]

Similar to CAC, a scoring system was developed by Mount Sinai to evaluate emphysema on LDCT scans [92]. The scores range from 0 to 3 describing no, mild, moderate, and severe emphysema, respectively, and are defined as follows:

- Mild: no discrete regions of hypoattenuation, but other parenchymal abnormalities suggesting the presence of emphysema.
- Moderate: discrete regions of hypoattenuation present and involved in less than half of the lung parenchyma.
- Severe: discrete regions of hypoattenuation present and involved in more than half of the lung parenchyma.

For the purposes of this research, the emphysema severity categories were grouped based on the presence of hypoattenuated regions (e.g., clear visible signs of emphysema) for a binary classification problem; a negative scan corresponded to no emphysema while a positive scan corresponded to mild/moderate/severe. The distribution of emphysema severity categories, dominant phenotypes, and class distribution can be found in Table 2.1. Note that 5 cases

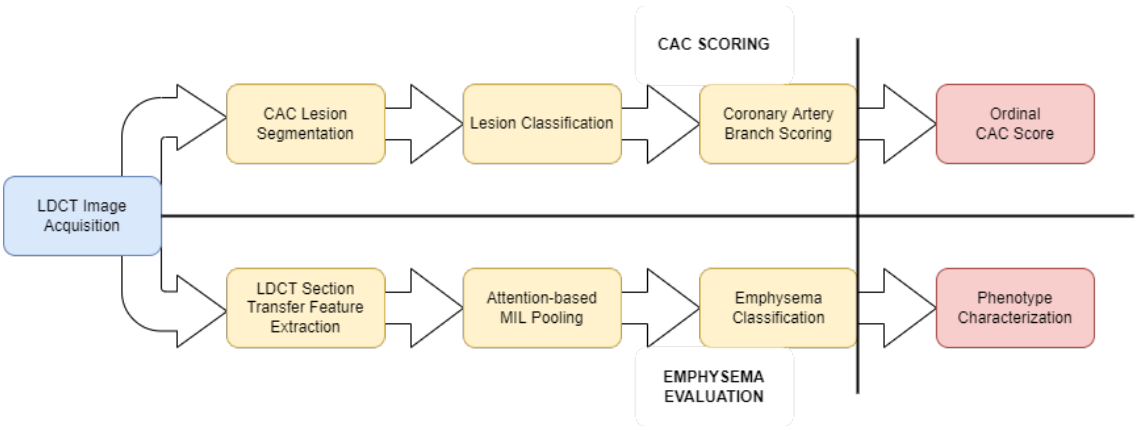

Figure 2.3: Proposed pipeline following LDCT acquisition. Note that the automatic evaluation is tailored to the specific condition based on desired output.

were excluded due to scan not encompassing the entire lung parenchyma, resulting in a total of 860 cases evaluated for the emphysema AI research.

In summary, the work completed in Chapters 3 and 4 fulfills the pipeline proposed in Fig. 2.3 that can be integrated into the lung screening workflow.

CHAPTER 3

AUTOMATIC SEGMENTATION AND CLASSIFICATION OF CORONARY ARTERY CALCIFICATIONS ON LOW-DOSE THORACIC CT

3.1 Coronary artery calcium

We build upon our knowledge of CAC, as summarized in Section 2.3, to develop machine learning methods for automatic assessment of CAC. As previously mentioned, there are several shared risk factors between lung cancer and severe CHD, the leading cause of death in the United States at an estimated 366,000 deaths per year and present in 18.2 million adults, thus it is fitting and appropriate to evaluate CAC at LDCT acquisition [72, 80, 83, 84]. However, a manual review of all LDCT scans for precise CAC score would be time-consuming and susceptible to reader variability errors, thus the goal of this study is the development of an automatic, rapid, and objective scoring system [93].

Several groups have previously produced machine learning models to evaluate CAC score on CT scans. Lessmann proposed a pair of deep networks with different receptive field sizes to identify potential CAC lesions and reduce false positive detections [94]. De Vos registered cardiac and chest CT scans to a 3D atlas to identify a field of view and relevant slices then used a CNN to identify CAC lesions, then produced a risk model for cardiovascular mortality [95, 96]. Zeleznik utilized a series of U-Nets to identify a field of view around the heart through automatic segmentation, then classified pixels as CAC or not [97]. Cano-Espinosa manually identified the heart in CT scans and used a 3D CNN to infer the Agatston CAC score [98]. Wang applied a threshold of 130 HU, as given by the Agatston score criteria, to identify all potential CAC candidates then classified each using a 3D ResNet architecture by location/branch as non-CAC, left anterior descending (LAD), circumflex (LCx), left main (LM), and right coronary (RCA) [99]. In this work, we expand upon prior

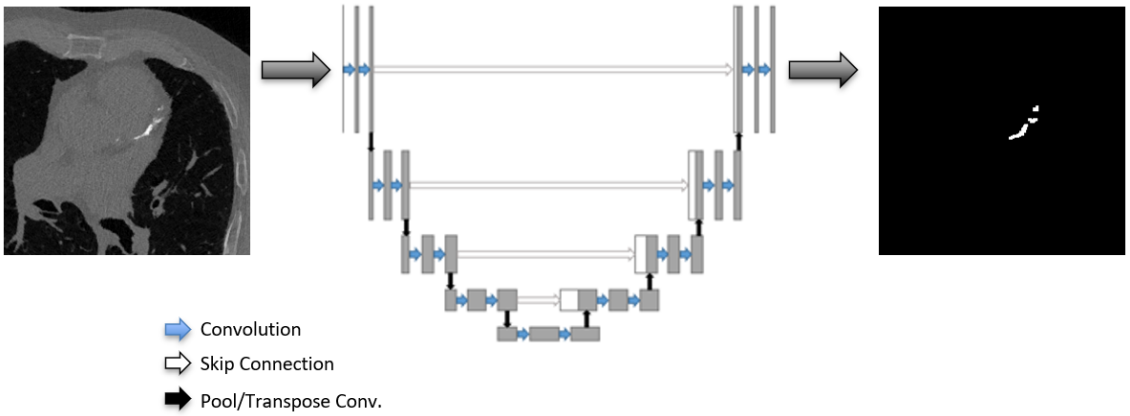

Figure 3.1: Standard U-Net architecture demonstrating an example segmentation of CAC within a CT section. The U-Net is composed of an encoding path, which utilizes convolutional and pooling layers to form an encoded representation of the input 2D image. The second half of the network then produces a binary segmentation map by decoding the representation into a probability map indicating if each pixel belongs to the desired structure/anatomy.

studies by performing segmentation and location-based lesion classification using a single U-Net based algorithm and providing CAC scores for an ordinal scoring scheme [100].

3.2 Revised U-Net Architecture

The standard U-Net architecture consists of an encoding path, which forms a latent representation of the input image, and a decoding path, which utilizes the latent representation to produce a binary segmentation (Figure 3.1) [101]. A qualitative evaluation of a multi-class U-Net for CAC lesion segmentation and classification found that the intermediate layers occasionally failed to maintain the long range spatial information needed for location-based classification through the late stages of the decoding path. Thus, proposed here is a U-Net variant called CACU-Net which preserves the information needed for CAC classification through an additional decoding branch with minimal additional training parameters.

The branch is composed of feature maps extracted from each decoding path level and combined through successive 2x2 kernel bilinear interpolations for dimensional matching

and channel-wise concatenation, thus carrying the long-range spatial information to the classification layer L2 (Fig 3.2) which identifies rough artery regions. Consequently, the reduced number of operations causes a tradeoff between the precise semantic information needed for accurate segmentation in this branch; however, this was maintained in the L1 segmentation layer which operates as a standard U-Net decoding path. The L2 layer classified between 5 location categories, LM, RCA, CFx, LAD, and None, with None indicating that the lesion was falsely identified as CAC. The L1 and L2 layer outputs were combined through elementwise multiplication of the L1 output with each channel of the L2 layer (e.g., arteries). The final artery classification decision was then taken as the channel in which the L2 layer provided the greatest probability. The full architecture is depicted in Fig 3.2, illustrating the two architecture branches.

3.3 Training, Testing, and Statistical Analyses

The two architecture branches were trained simultaneously with loss function L = L1 + L2 where L1 was the binary cross-entropy loss calculated on the segmentation from the main branch and L2 was the categorical cross-entropy loss calculated between the predicted classification map and the reference standard artery segmentations. In this way, the L1 classification layer was optimized for sensitivity and segmentation accuracy while the L2 classification layer was optimized for artery classification. There are many potential sources of false positive CAC detections in LDCT scans and unnecessary patient recall is a key problem, thus, we utilized ROC analysis to first determine an optimal LDCT scan level CAC severity decision threshold for each artery [102]. Performance was characterized by the area under the ROC curve (AUC), with the average area of calcium per CT scan slice serving as a pseudo-volume score due to varying patient sizes and slice thicknesses.

Five models were trained with different randomly produced training, validation, and testing sets consisting of 80%, 10%, and 10% of the available cases, respectively, and AUC

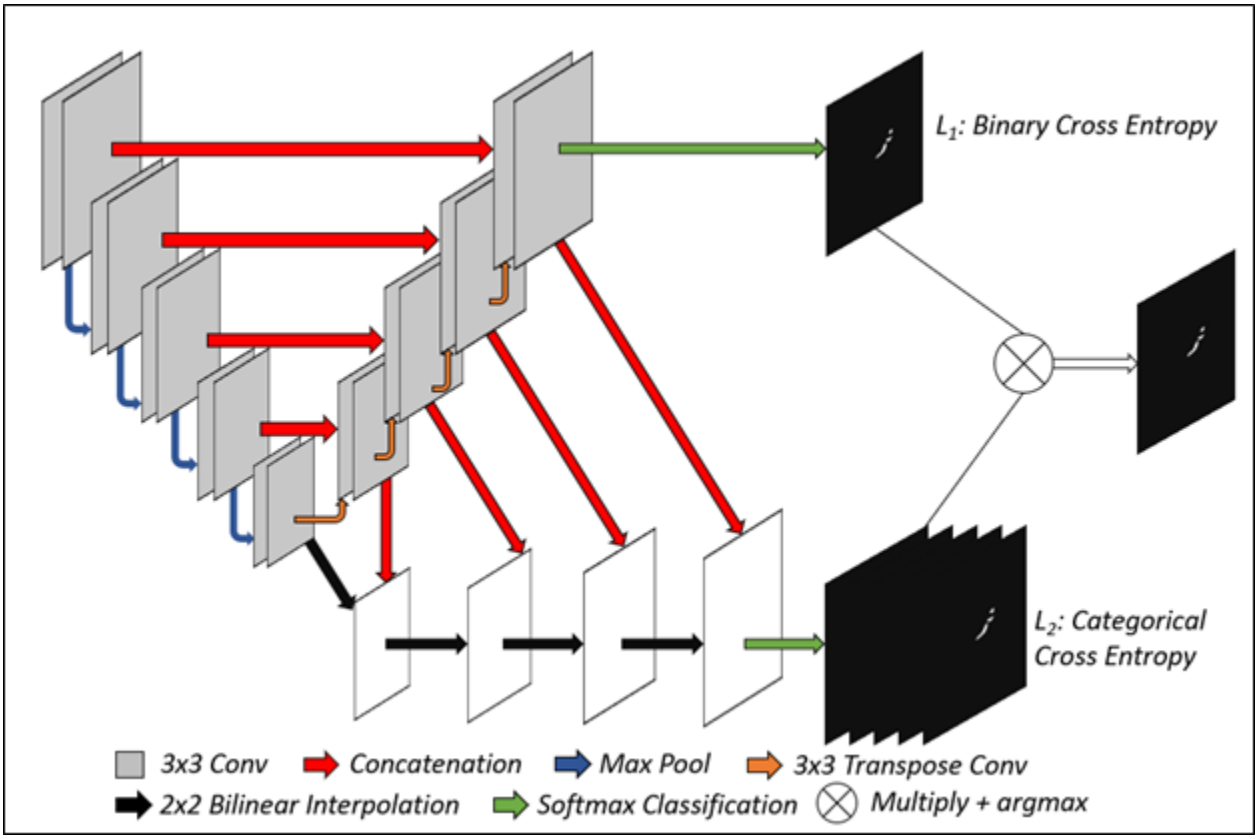

Figure 3.2: Schematic of the proposed deep network architecture, CACU-Net. Note that the softmax classification layers labelled with loss functions L1 and L2 attempt to match target outputs, with L1 aiming to segment all lesions and L2 aiming to segment lesions in each artery branch with different channels represents the different artery branches. A threshold was then applied to the L1 output to provide a binary segmentation which was then multiplied in an elementwise manner with each of the artery class probability maps produced by the L2 prediction layer. Artery class was then determined on a per-pixel basis by the maximum artery class probability for each pixel (LMA, RCA, CFX, LAD, and None).

variance was obtained across the five models. AUCs were compared through the DeLong test on each of the five training passes with the median p-value serving as the metric for significance [103]. The network was trained with Adam optimization with learning rate 0.01, batch size 8, and stopping patience of 5 training passes of no improvement. Note that the architecture chosen for this study is 2D rather than 3D, since, although 3D models have increased number of parameters with the potential for improved performances, they also require substantially more data for training.

All machine learning analysis was performed in Python 3.0 using Keras 2.7 with Tensorflow backend while image analysis and reference standard segmentations for U-Net training were provided through MATLAB version 2020b with the MATLAB Image Labeler. Statistical tests also included the Stuart Maxwell test [104].

3.4 ROC Analysis and Severity Evaluation

In the task of determining whether a LDCT scan presents with CAC in any artery, the CACU-Net performed with an area under the ROC curve of 0.94 (0.03), which demonstrated comparable performance to the standard U-Net algorithm performance of AUC 0.94 (0.03) and outperformed the multi-class U-Net variant with AUC 0.90 (0.16). Further, CACU-Net also showed significant improvements for the task of detecting CAC in the LM, CFx, and RCA, including improved consistency across all arteries, and failed to demonstrate a significant difference for the LAD. ROC performance and comparisons are displayed in Table 3.1.

The confusion matrices in Figures 3.3 and 3.4 compare the ability of the three U-Net variants in the tasks of stratifying LDCT scans and coronary artery CAC scores into clinically relevant categories. At the coronary artery level, CACU-Net correctly predicted the ordinal class for 67% of scans for the LAD, 70% for CFx, 74% for RCA, and 80% for the LM. These results contrast with the ROC performance because the CACU-Net tended to underestimate

Table 3.1: Scan and Artery Performance AUCs (standard deviation). Bold indicates statistical significance.

Model	Scan	RCA	LMA	LAD	CFX
Standard U-Net	0.94 (0.03)	-	-	-	-
Multi-class U-Net	0.90 (0.16)	0.86 (0.13)	0.62 (0.19)	0.94 (0.05)	0.81 (0.18)
CACU-Net	**0.94 (0.03)**	**0.87 (0.04)**	**0.77 (0.06)**	0.92 (0.04)	**0.86 (0.03)**

the ordinal score, thus the arteries that had generally lower, less variable scores provided more accurate severity evaluations. However, the Stuart Maxwell test for each artery did demonstrate a significant difference between CACU-Net and radiologist scores ($p<0.05$) for each of the four coronary artery branches, suggesting further improvement is needed to match radiologist performance.

3.5 Evaluation of Severe Misclassification

We define the classification error as

$$\epsilon_{ij} = |y_{ij} - f(x_{ij})|$$

where y_{ij} is the radiologist-determined ordinal score for the jth artery of the ith case and $f(x_{ij})$ indicates the predicted ordinal score. Severe misclassification at the scan level is defined where $\sum_j \epsilon_{ij} \geq 3$, which can be acquired either through a combination of artery level misclassifications, of which there can be major ($\epsilon_{ij} \geq 2$) or minor ($\epsilon_{ij} = 1$) misclassifications. Manual review of cases revealed that 11.0% of cases suffered severe misclassification, with 5.0% underscored and 6.0% overscored.

Of the 5% of CT scans (36 scans) that were underscored, 60.0% were caused by major misclassification in at least one artery, with 28.5% of those also experiencing more than one

a)

LAD

Predicted Severity \ Ground Truth Severity	0	1	2	3
0	256	116	5	0
1	8	142	30	1
2	1	33	47	20
3	1	13	28	19

CFx

Predicted Severity \ Ground Truth Severity	0	1	2	3
0	411	67	7	2
1	50	84	26	3
2	6	11	9	3
3	5	10	9	17

LM

Predicted Severity \ Ground Truth Severity	0	1	2	3
0	505	49	6	3
1	43	18	9	7
2	16	6	6	3
3	23	7	12	9

RCA

Predicted Severity \ Ground Truth Severity	0	1	2	3
0	430	82	6	2
1	29	83	20	5
2	2	7	8	9
3	4	12	10	11

Figure 3.3: Confusion matrices for the individual coronary arteries comparing predicted and reference standard ordinal scores for each of the four main coronary artery branches.

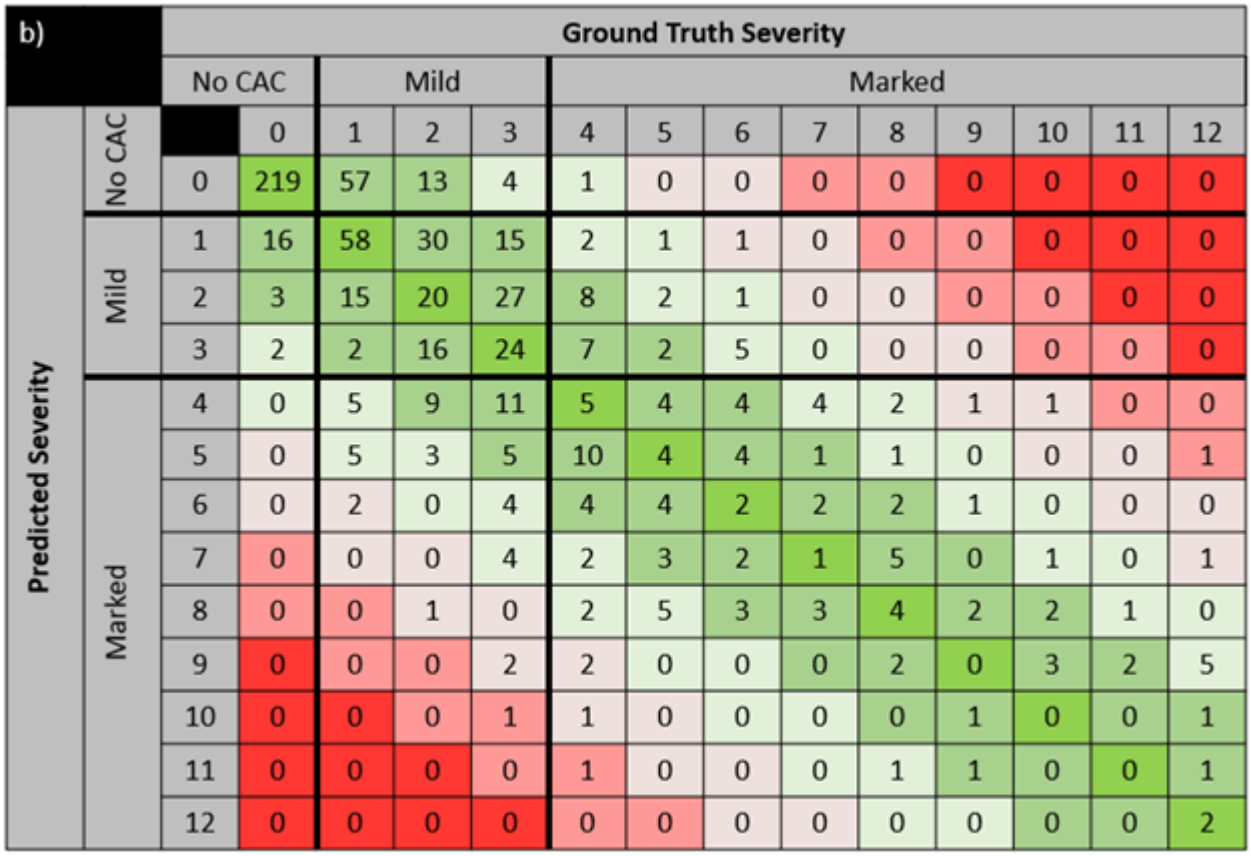

Figure 3.4: Confusion matrices for the scan level MSOS comparison with clinically relevant partitions of no (score: 0), mild (score: 1-3), and marked (score: 4-12) severity. Note that this is the confusion matrix of only one of the five models, but is representative of the results from all five.

major artery misclassification. The remaining 40% were caused by a combination of minor misclassifications.

Of the 6% of CT scans (43 scans) that were overscored, 86.0% tended to be more impacted by major misclassifications in at least one artery, with the remaining 14.0% caused by a combination of minor misclassifications. 32.5% of the overscored cases contained multiple severe artery misclassifications.

3.6 Example Segmentations

The example images in Fig. 3.5 provide insight into successful and failed model use cases. Fig. 4a-d depict successful segmentation and classification performance for each artery. Fig. 3.5e-h demonstrate vulnerability to motion artifacts through severe oversegmentation (Fig. 3.5e) or misclassification (Fig. 3.5g,h). There were also occasional instances of completely missed CAC lesion segmentations as in Fig. 3.5f, which marks a missed RCA lesion, potentially due to close proximity and appearance to the ribs.

3.7 Discussion of CACU-Net and Future Directions

The novel architecture additions proposed in this study showed negligible detrimental effects on scan level performance compared to the standard U-Net with both demonstrating AUCs of 0.94 (0.03) despite the additional information extracted in the CACU-Net. The proposed model also performed more consistently across random testing sets than a more standard multi-class U-Net variant for each artery. Further, ROC analysis showed that the artery classifications in order of increasing performance were LMA, CFX, RCA, and LAD.

Based on the ROC and confusion matrix analysis, there is strong potential for clinical utilization of the proposed architecture for automatic calculation of the ordinal CAC severity score on LDCT scans. The additional information acquired through the classification branch

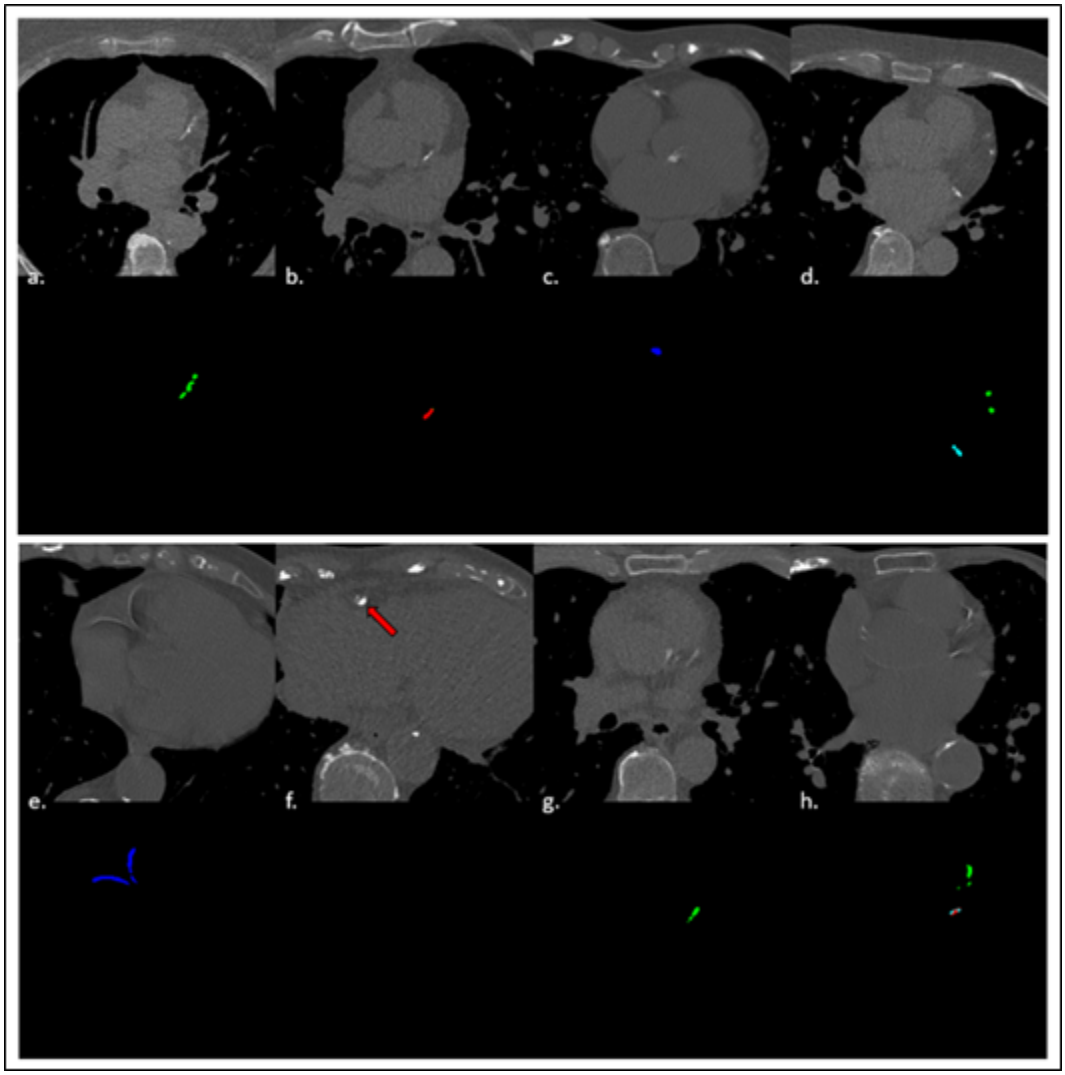

Figure 3.5: Cropped example images and their corresponding automatically produced segmentations. Colors indicate artery classification, with green=LAD, red=LMA, blue=RCA, cyan=CFX. Image pairs (a-d) display examples of successful identification and classification for each artery while image pairs (e-h) demonstrate common cases of partial or complete failure in either segmentation (e-f) or classification (g-h). The red arrow in (f) marks an RCA lesion that was completely missed.

of the model holds significant clinical value as recent studies have suggested more effective cardiovascular risk metrics by including location information to more accurately identify potentially dangerous lesions [96, 105–107]. For example, de Vos found that TAC deposits were the most significant predictor of cardiovascular risk [96]. While the currently proposed model does not account for TAC due to lack of reference standard or inclusion in the ordinal scoring system, this could be included as an additional class in the L2 classification layer with retraining.

The confusion matrices support the ROC conclusions, with the LMA demonstrating worst performance followed by RCA, CFX, and LAD in order of improving performance. The model had a slight tendency to overpredict severity; qualitative case review demonstrated that these were often due to motion artifacts, which has been explored in prior literature. The qualitative review also found few cases where mitral annular calcifications were misclassified as CFX lesions. This could be due to the inclusion of the long-range spatial information in CACU-Net classification, but further investigation is needed to confirm this. The major misclassifications were often caused by severe LMA misclassification, which may be explained by two factors: 1) the LMA had the lowest CAC prevalence of all lesions in the dataset, which is potentially desirable because it is representative of a real clinical population but may have been suboptimal for deep network training purposes, and 2) it is difficult to distinguish LMA lesions from LAD and CFX lesions in certain situations. This supports the decision in other studies to combine LAD and LMA lesions in a single class.

This study was limited by a lack of comparison to similar segmentation approaches which is partially due to the comparison of different datasets, which may be inappropriate, and more so that this study aimed to calculate the ordinal score instead of the Agatston score. While the scores have been shown to correlate well, Agatston calculations were not performed for this study due to a lack of reference standard data, despite the proposed algorithm's capability to produce the Agatston score [86, 100]. Further, the reference lesion segmentations

used for model training were subject to potential errors due to motion artifacts and other sources of variability; further investigation is required to determine if other methods, such as the CycleGAN based method proposed by van Velzen, are more effective in addressing these issues [108]. Finally, as noted, this study only classified CAC lesions as opposed to all potential sources of calcium within the LDCT scan. While additional classification options could easily be added to the proposed model, this would require model re-training.

In summary, we developed a novel U-Net variant that segments and classifies CAC lesions with efficient, effective architectural additions. The proposed model was able to identify scans which contained clinically relevant CAC with high performance in a population of LDCT cases, demonstrating strong clinical potential for this method. Further, this approach could be used for the development of improved cardiovascular risk metrics, specifically those that incorporate CAC volume, density, and location, and may aid in reducing variability during radiologist reporting.

CHAPTER 4

EMPHYSEMA CHARACTERIZATION THROUGH MULTIPLE INSTANCE TRANSFER LEARNING IN LUNG SCREENING CT SCANS

4.1 Multiple Instance Learning

In this chapter, we attempt to characterize emphysema to complete the LDCT evaluation pipeline. As discussed in Section 2.4, this problem significantly differs from CAC scoring, requiring characterization of regions including all lung tissue rather than identification of individual lesions.

Multiple instance learning (MIL) is a deep learning scheme commonly used in digital pathology that utilizes weak annotations to train models by evaluation of instances (e.g., CT sections) to form a collective classification decision of a bag (e.g., CT scan) [109]. Wang discussed key MIL schemes, mi-Net and MI-Net, which classify scans based on individual instance classifications and pooled instance representations, respectively [110]. Ilse improved MIL schemes through attention-based multiple instance learning, which utilizes attention mechanisms to identify and more heavily weight key instances of whole slide images for cancer detection [5].

Deep learning, including MIL schemes, have been utilized to automate emphysema evaluation in standard diagnostic and lung screening CT scans. Humphries utilized a convolutional neural network and long short-term memory architecture to classify visual emphysema pattern on CT and Oh used the same model to compare visual emphysema progression with functional impairment and mortality [111, 112]. Negahdar automatically segmented lung volumes on chest CT and classified patches of lung tissue based on visual emphysema pattern to quantify severity [113]. Chepylgina and Orting utilized human-engineered features based on histogram features acquired from filtered lung ROIs in a multiple instance learning

scheme to characterize COPD and emphysema, respectively, in low-dose CT scans [114, 115]. Tennakoon expanded their work to incorporate deep MIL on 3D LDCT patches to classify emphysema presence [116].

In our work, we utilize deep MIL with transfer learning and attention-based pooling (Transfer AMIL) to evaluate emphysema in LDCT scans and compare performances in classification of disease.

4.2 Transfer Multiple Instance Learning

Typically, MIL is posed as a binary classification problem in which the data are composed into bags $X_i = \{x_{i1},\ x_{i2},\ ...,\ x_{iN}\}$ each of which is composed of N instances x_{ij} [5, 110]. The corresponding instance truths $y_{ij} \in \{0, 1\}$ are unknown, but the bag truth is determined from the instance truths by the binary decision rule.

$$Y_i = \begin{cases} 0, & \text{iff} \quad \sum_j^N y_{ij} = 0 \\ 1, & \text{otherwise} \end{cases} \tag{4.1}$$

MIL can be broken down into three key steps as: 1) extraction of instance representations, 2) transformation from instance representations to bag representation through MIL pooling, and 3) classification of bag representation for clinically relevant decision [117]. In all, the process is described by

$$\hat{Y_i} = g(\mathbf{P} f(\mathbf{X_i}))$$

where $\hat{Y_i}$ is the predicted bag label, $\mathbf{X_i}$ is the set of input CT sections (images) that are transformed to instance representations via f, pooled via matrix $\mathbf{P}$, and transformed to a bag prediction via g [117].

In our study, instant representations $f(\mathbf{X_i})$ of CT sections are acquired through transfer learning from a pre-trained VGG19 architecture [118]. Transfer learning utilizes large models

with deep, hierarchical features after pre-training for a similar task, in this case image classification but on the ImageNet database set of natural objects [13, 14, 119]. In situations where little training data are available, transfer learning allows for the extraction of more complex, rich data representations than can be achieved by training a model from scratch. In this study, we utilized a VGG-19 architecture to extract quantitative features similar to the scheme proposed by Antropova [4].

The instance representations were then input to two fully connected layers with ReLU activation with a dropout rate of 0.5.

Attention mechanisms have been widely utilized in deep learning to both improve performance and provide interpretability of model predictions [120]. In our study, the pooling matrix $\mathbf{P}$ was constructed through the MIL attention mechanism in which a bag representation was acquired through a weighted average of instance representations:

$$\mathbf{z} = \sum_{n=1}^{N} a_n \mathbf{x_n} \tag{4.2}$$

$$a_n = \frac{\exp(\mathbf{w^T} \tanh \mathbf{V x_n^T})}{\sum_{j=1}^{N} \exp(\mathbf{w^T} \tanh \mathbf{V x_j^T})} \tag{4.3}$$

for learned parameters $\mathbf{w} \in \mathbb{R}^{128}$ and $\mathbf{V} \in \mathbb{R}^{128x512}$ with N input instances $\mathbf{x_n^T}$ with dimension 512 and hidden dimension 128. The attention weights also provided interpretable output inherent to the decision task in the form of influential instances (i.e., slices), which were evaluated separately for model validation and interpretability.

The attention weights for different scan classes (dominant emphysema phenotypes of centrilobular, panlobular, and paraseptal) were evaluated by scaling attention weights for a given scan to the range [0,1] and plotting as a function of the axial depth to determine regions of high and low influence. Influence was quantified by three metrics: 1) depth maximum

attention of fit curve, 2) weighted average of slice depths weighted by attention, and 3) range of fit curve attention values. The full workflow of Transfer AMIL is provided in Fig 4.1.

The attention weights for different scan classes (dominant emphysema phenotypes of centrilobular, panlobular, and paraseptal) were evaluated by scaling attention weights for a given scan to the range [0,1] and plotting as a function of the axial depth to determine regions of high and low influence. Influence was quantified by three metrics: 1) depth maximum attention of fit curve, 2) weighted average of slice depths weighted by attention, and 3) range of fit curve attention values. The full workflow of Transfer AMIL is provided in Fig 4.1.

4.3 Training, Testing, and Statistical Analyses

All models were trained in Keras (2.2.4) with Tensorflow backend (2.2.0) in Python (3.7) and optimized by binary cross entropy loss calculated for bag predictions. Adam optimization was utilized with parameters $\beta_1 = 0.9$ and $\beta_2 = 0.99$ and initial learning rate of 0.0001. Early stopping was initiated if the validation loss did not improve after 7 iterations. All learned parameters were initialized by sampling a normal distribution.

Models were trained through 5-fold cross validation by case with 60%, 20%, and 20% of the available cases serving for training, validation, and testing, respectively, in each evaluation fold. The mean and variance of the area under the ROC curve (AUC) were obtained across the five models. AUCs were compared through the DeLong test on each of the five training passes with the median p-value serving as the metric for significance [103].

We compared Transfer AMIL to other approaches which required only scan annotations. A 3D CNN classifier was trained by interpolating to a fixed input size of 128 slices and scan presence of emphysema serving as binary class. Additionally, a standard 2D classifier was trained by assigning the scan class label to all slices within the scan regardless of emphysema presence within that slice; this caused noisy labels during training, particularly with many false positive slices for severe emphysema cases.

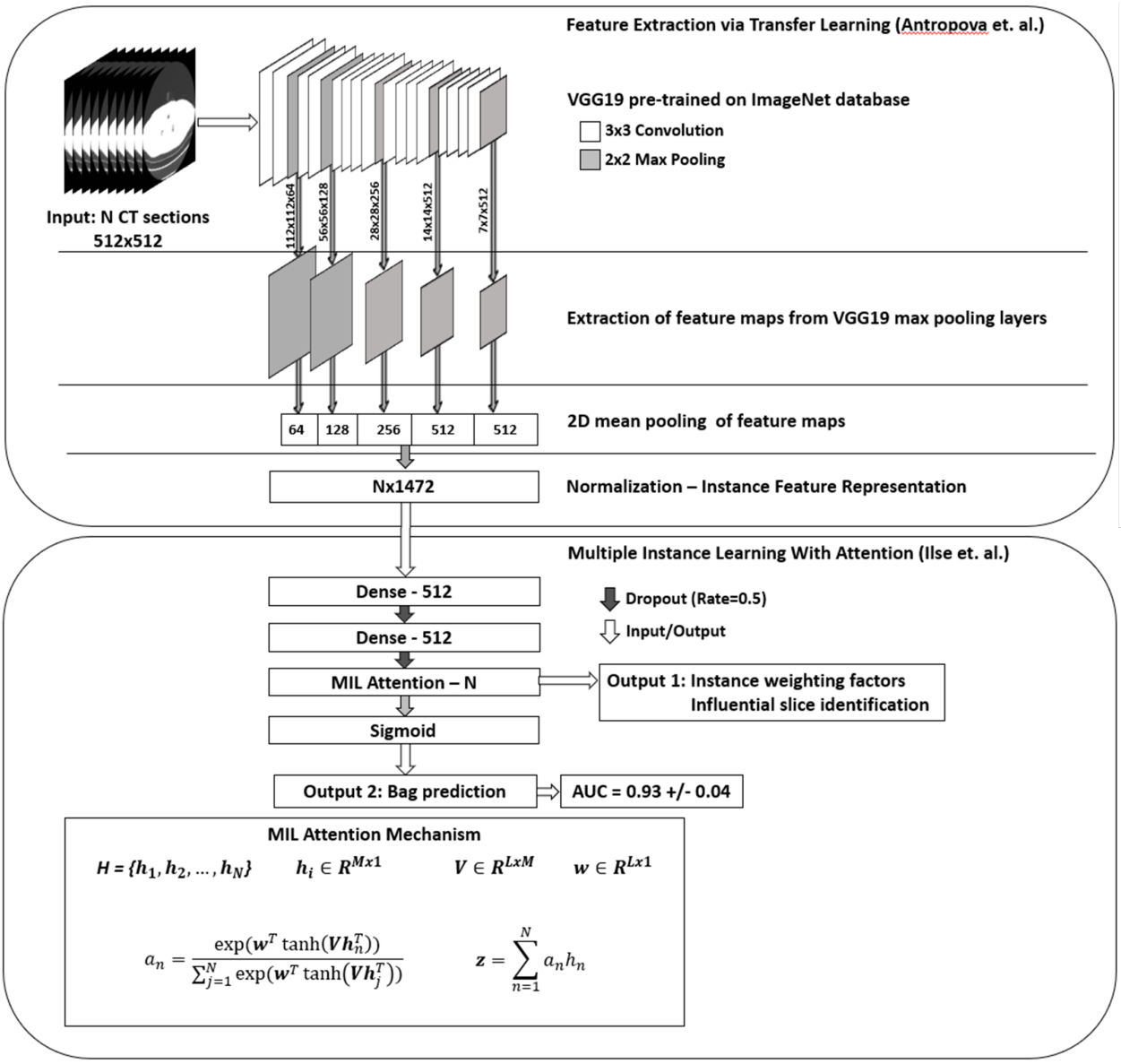

Figure 4.1: Model workflow of the Transfer AMIL approach. This includes feature extraction of CT images through an ImageNet pre-trained model based on methods developed by Antropova et. al. followed by attention-based MIL pooling based on methods developed by Ilse et. al. [4, 5]. Two outputs are generated for each LDCT scan input, the attention weights which identify influential slices for the classification task and the scan prediction for the presence of emphysema.

Table 4.1: Emphysema classification assessment

Algorithm	AUC from ROC Analysis	Human-Engineered Features	Deep CNN Features	Interpretable	Transfer Learning
Transfer AMIL	**0.94 +/- 0.04**		**X**	**X**	**X**
Noisy 2D Classifier	0.85 +/- 0.06		X		
Fully 3D Classifier	0.58 +/- 0.16		X		
AMIL	0.69 +/- 0.05		X	X	
Mean Pooling	0.90 +/- 0.02		X		X
Max Pooling	0.88 +/- 0.02		X		X
Cheplygina	0.78 +/- 0.04	X		X	
Orting	0.88 +/- -.—-	X			
Tennakoon	0.95 +/- -.—-		X		

4.4 Binary Classification Performance

In the task of determining if a CT scan presented with emphysema or not, the Transfer AMIL approach yielded an area under the ROC curve of 0.94 +/- 0.04, which was a statistically significant improvement compared to other methods evaluated in our study following the DeLong Test with correction for multiple comparisons (Table 4.1). Transfer AMIL performed better than or similar to other published work, including shallow, human-engineered MIL methods, as well as other deep MIL approaches, although it is important to note that others' evaluations were on different datasets.

4.5 Attention Weight Interpretability Analysis

Attention weight curves were calculated to demonstrate the influence of disease type localized throughout the lung. The attention weights demonstrated a stronger influence for slices in the upper lung in all scan classes, indicating that the model prioritized upper lobe information

Table 4.2: Quantitative attention weights from Figure 4.2

Scans Evaluated	Maximum Attention Lung Depth (%)	Weighted Average of Lung Depth (%)	Range of Attention Values (%)
Positive Scans	15.6	38.5	33.6
Negative Scans	32.9	38.7	47.2
Positive: Centrilobular	19.3	39.1	34.8
Positive: Panlobular	17.2	46.2	20.0
Positive: Paraseptal	12.8	37.6	24.1

(Table 4.2, Figure 4.2). This agrees with published literature trends that note an upper lobe predominance for emphysema, particularly centrilobular, the most common phenotype in this dataset [88–91]. Recall, influence is quantified by three metrics: 1) depth maximum attention of fit curve, 2) weighted average of slice depths weighted by attention, and 3) range of fit curve attention values.

By phenotype, the centrilobular and paraseptal attention average depths (39.1%, 37.6%) aligned with expected upper lobe predominance compared to panlobular (46.2%). Further, the panlobular scans tended to more heavily influence slices throughout the lung range as demonstrated by the reduced range of attention values (20.0%) compared to the other phenotypes (34.8%, 34.1%). Note that any given scan did not necessarily present with only one phenotype; for example, the scans labeled panlobular-dominant may also present with other phenotypes. This and the model's learned predisposition to more highly weight the upper lobe slices (as conveyed by the quantification of negative scan attention) may account for the relative importance of the upper lobes even in the panlobular-dominant scans.

The top-k influential slices according to attention weights were evaluated to determine which CT imaging features drew the most attention and to identify potential sources of misclassification. The prevalence of image features that were present in the top-k selected attention weighted slices are shown in Fig 4.3. Different features were likely to have different prevalence within each scan (e.g., nodules were local abnormalities while architectural

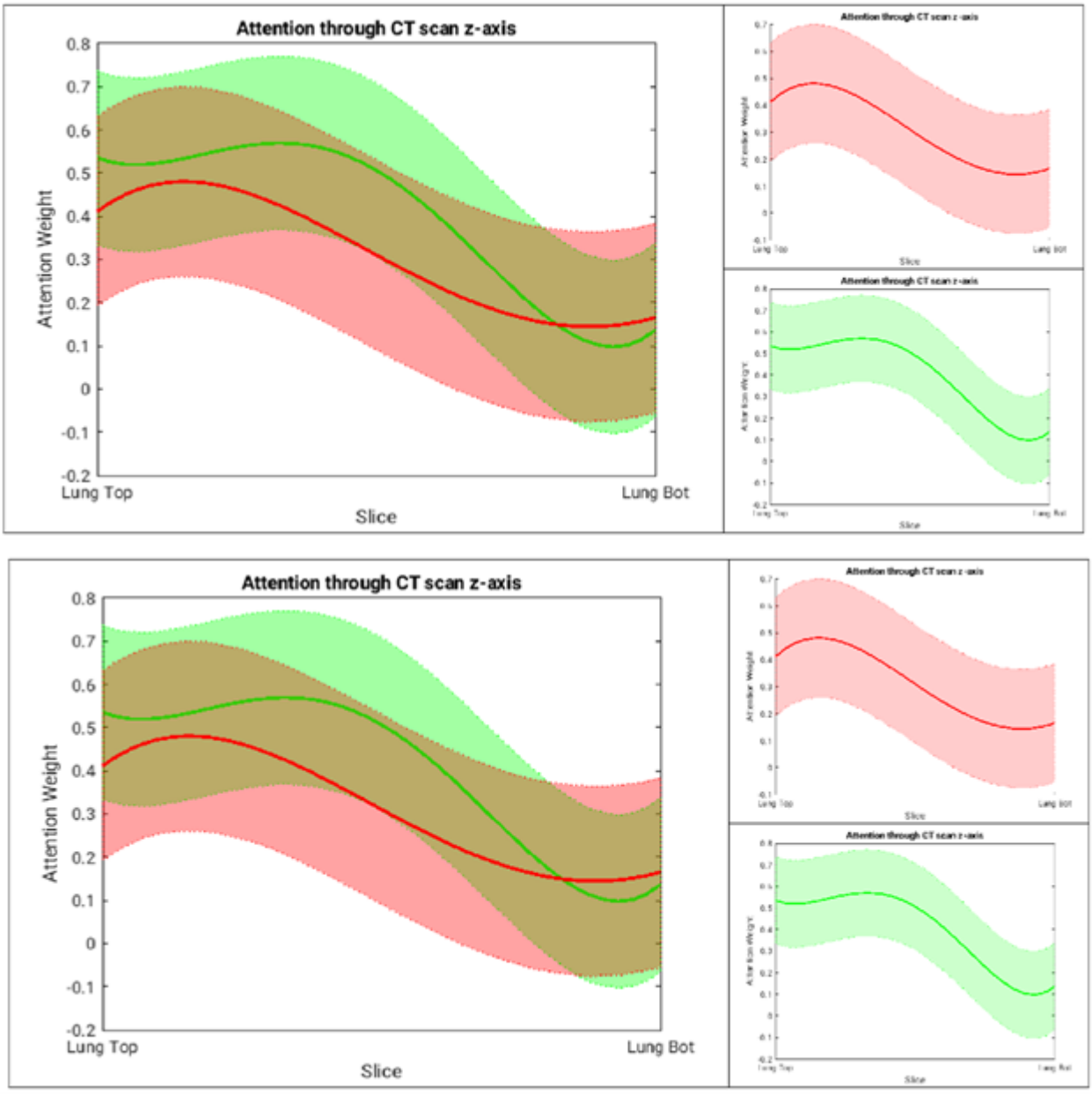

Figure 4.2: Attention weight curves illustrating the fit of attention weights from CT slices as a function of height in the lungs for (top) positive (red) and negative (green) LDCT scans and (bottom) for different dominant phenotypes of emphysema: centrilobular (blue), panlobular (pink), and paraseptal (turquoise). Since patients' CT scans have variable number of slices covering the lung region, in these plots, the range has been normalized to fit between Lung Top and Lung Bottom.

distortions were generally more widespread structural changes), thus the prevalence of each imaging feature as identified by a radiologist is presented for comparison. Bronchial disease and architectural distortions demonstrated the largest change in importance for the top-k attended slices compared to the human reader with changes of 23.7% +/- 0.03% and 22.7% +/- 0.01%, respectively. The prevalence of each feature did not significantly change when including more slices in the attention analysis; however, bronchial disease and architectural distortion features were attended to much more frequently than their frequency in entire scans while the opposite occurred for ground glass opacities. This may suggest that the model was balanced between identifying features indicative of emphysema presence, such as regions of hypoattenuation and structural changes, while maintaining a general representation of the entire CT scan.

4.6 Discussion and Future Directions

In this study, we present a novel CT slice-based Transfer AMIL approach for evaluating emphysema on LDCT scans acquired for lung screening. The model provides strong classification performance compared to models with similar label constraints, including models evaluated for this study and those published in the literature. The attention module also provides interpretable information for verifying model performance by identifying slices that were most influential to the classification decision. Indeed, the attention weight trends for different subsets of the LDCT scans agreed with expectations in terms of the most likely regions to find emphysema, including when different classes of emphysema were dominant. A further investigation into the attention weights also revealed which CT image features were most useful for the model prediction and may provide insight into what potential cases will be problematic for automatic evaluation, particularly considering the lung screening population.

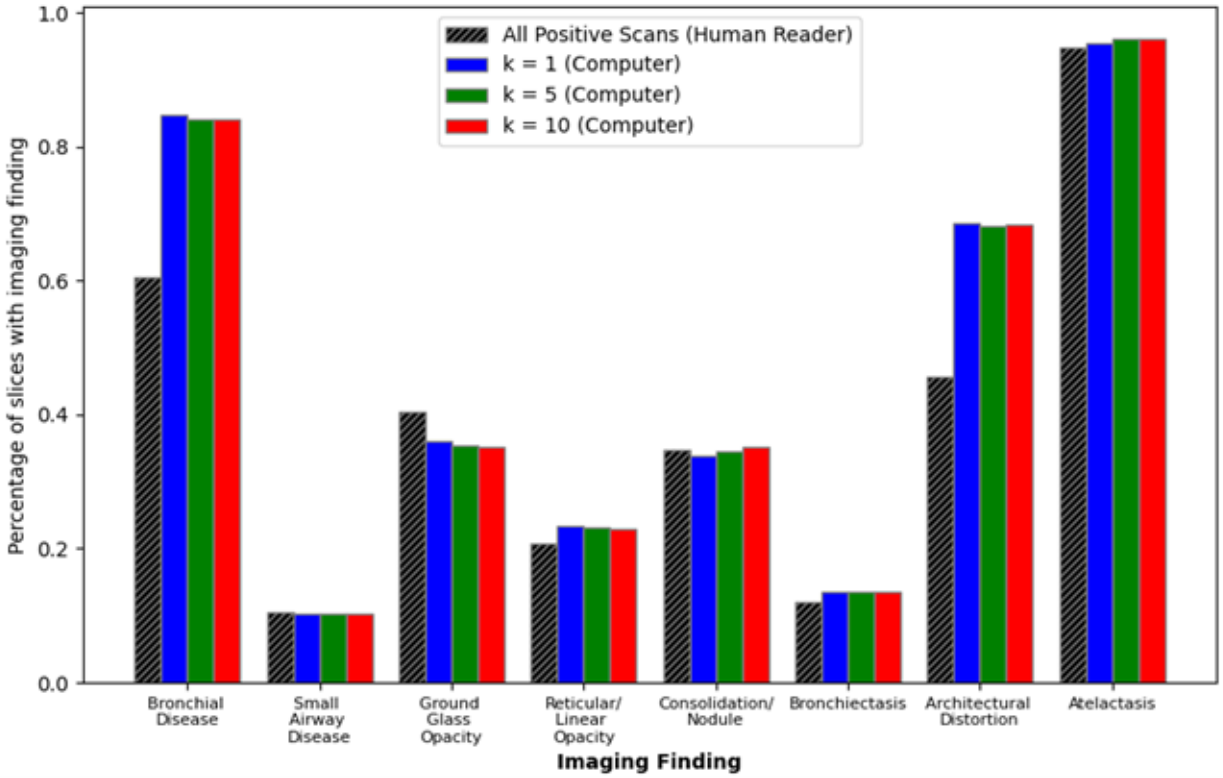

Figure 4.3: Evaluation of common thoracic imaging features. The prevalence of each feature within the entire CT scan as identified by a radiologist and when selected by the top-k attention weighted slices. Note key differences between whole slice prevalence and selected prevalence: bronchial disease and architectural distortions were more heavily weighted while ground glass opacities are diminished. Further, the consistent representation across the top-k slices for different k demonstrates the model's tendency to more heavily weight slices with similar extracted representations.

Importantly, the developed model requires a relatively small amount of computing power compared to other modern deep learning computer vision tasks. The AUC performance achieved by the Transfer AMIL was either comparable or outperformed other models in this study, including standard 2D classification models with noisy labels and 3D image classifiers. Note that this performance may only hold true for the data available for this study; a large amount of data would likely improve the non-transfer learning models more than transferred models because during training the number of training images approaches the number of trainable parameters.

The pre-trained VGG19 feature extraction model parameters (20M) were fixed from the pre-training task with no additional training; additional training would further underdetermine the model considering the limited dataset of 860 scans. With the feature extractor fixed, the additional fully connected layers and attention module require only 1.15M trainable parameters; still an underdetermined system, but at a greatly reduced risk of overfitting. While larger standard architectures such as ResNet50 and DenseNet121 could be utilized for feature extraction, this study demonstrates that even the use of smaller, less complex models can achieve competitive performance. Note, these architectures were evaluated but no performance gain was observed thus the least computationally expensive model was utilized. Further, the reduced model capacity and use of transfer learning with a common architecture encourage wider implementation of this technique because the compute power needed to run the model is generally attainable by today's standards and feature extractor does not require local training.

While the attention module interpretable output validates model performance compared to other published studies, it also encourages clinical implementation as the attention weights can be added as an optional part of the lung screening workflow for a radiologist to further investigate the classification decision, specifically by review of the slices influential to the classification decision. This review process can lead to radiologist trust and understanding

of clinical implementation of the algorithm and has the potential to improve clinical workflow in terms of both reading time and performance, although this would require a prospective reader study to confirm.

The theme of improved performance also aligns with the attention module's ability to identify which cases may be problematic for classification. For example, the model tended to more heavily weight slices with bronchial disease and architectural distortions, which are nonspecific to emphysema patients, and which often appear similar to typical presentation of emphysema (e.g., regions of hypoattenuation and structural changes). This also suggests that patients with these presentations caused by non-emphysematous conditions may be difficult for the model to classify.

Future work should prospectively utilize this model in a reader study to evaluate its impact on radiologist performance and radiological workflow as well as include images acquired from multiple institutions to assess model generalizability. This is especially important as the data in this study were limited (single institution, single scanner manufacturer, limited N). Further, this study only evaluated binary classification decisions and does not consider relationships between slices when calculating attention weights; multi-class variants of MIL as well as more complex attention-based pooling functions. Despite these limitations, the Transfer AMIL method achieved strong performance as determined from ROC analysis and the attention weight investigations performed in this study demonstrated strong potential for clinical implementation.

CHAPTER 5

PROGNOSIS AND TREATMENT RECOMMENDATION FOR COVID-19 PATIENTS THROUGH DEEP LEARNING ON THORACIC CT

In the final aim of this dissertation research, we veer from the LDCT domain to evaluate CT scans of patients who have been infected with SARS-CoV-2, the virus causing COVID-19. In this chapter, we begin with a discussion of the epidemiological impact of COVID-19 and how medical imaging has been utilized to evaluate COVID-19 patients, particularly during the recent pandemic. We then present two separate but related studies: 1) our preliminary study which utilized a rudimentary MIL approach on a limited dataset from early in the pandemic, and 2) our more recent study that expanded the MIL approaches related to COVID-19 on a significantly larger, more representative dataset. The deep network models developed in this aim have the potential to improve COVID-19 patient management, through assessment of disease severity and prediction of outcomes for both hospital resource management and treatment options.

5.1 COVID-19 and Medical Imaging

The recent outbreak of the 2019 novel coronavirus has disrupted the global economy, exhausted medical resources, and adversely affected millions of individuals [121–123]. The associated disease (COVID-19) typically manifests through pulmonary dysfunction, including development of acute respiratory distress syndrome through COVID-19 pneumonia [124]. Steroid administration has been widely implemented by clinicians to treat severe cases of COVID-19 despite the many side effects that have been recognized [125, 126]. In particular, methylprednisolone and dexamethasone are common steroids used for COVID-19 treatment due to their demonstrated impact in treating inflammatory symptoms in other respiratory

infections [127, 128]. However, patient reaction to steroid administration is variable, depending on many factors including patient age, smoking history, and other comorbidities. Thoracic imaging through chest CT is used clinically to aid in differential diagnoses, monitor disease progression/severity, and, in the case of steroid administration, inform treatment regimen, which is especially critical for COVID-19 due to the significant burden that this disease places on medical resources. Deep transfer learning methods may have a role in identifying the amount and type of medical resources that will be needed throughout patient hospitalization.

The primary finding of COVID-19 patients on CT scans is peripheral and patchy or nodular/mass-like ground-glass opacities (GGO), typically presenting bilaterally with a predominance for lower lung lobes [129, 130]. The typical pattern is often reminiscent of organizing pneumonia. As the disease progresses to a severe state, GGO is observed more centrally, with infiltration and consolidation [129]. The visualization of COVID-19 through CT is strongly dependent on the amount of time between virus contraction and scan acquisition, potentially causing inaccurate diagnosis during image reading [131]. However, CT has been used by clinicians as the most effective way to visualize the progress of treatment for many pulmonary diseases, including lung cancer and pneumonia, by assessing changes in diseased tissue size, shape, and density. However, these evaluations are often qualitative and subjective, leading to inconsistent judgments and, potentially, detrimental consequences in patient care. Exploring quantitative metrics such as volume and density has shown improved evaluation accuracy, however these measurements depend on accurate, consistent delineation of the diseased tissue which requires consensus from radiologists to draw, reconcile, and prioritize their delineation. Deep learning has the potential to overcome these difficulties and provide quantitative assessments of disease progression.

5.1.1 COVID-19 Databases

Two datasets were collected from COVID-19 infected populations. The first dataset, COVID-Set1 (CS1), was acquired in the early days of the pandemic, February 2020-March 2020, prior to the discovery of effective treatment methods, and thus this set of 41 patients all reached severe disease stage and were referred to the Renmin Hospital of Wuhan University for treatment, including the use of ventilators, antivirals, and steroids. In particular, the decision to administer steroids was reached based on a combination of symptom severity and disease presentation on CT imaging, with 27 of the 41 patients determined by an expert intensivist as severe enough to necessitate steroid administration. For each of these patients, multiple CT scans were acquired throughout their course of treatment to monitor response and disease progression. However, the cases in this cohort were limited; i.e., all of the patients were responsive to administered treatments, thus none of them died. Further, as previously discussed in Chapter 2, deep learning models perform at their best when large amounts of data are provided, thus the total number of 41 patients greatly exposed any deep learning training to model overfitting and will be revisited in Section 2.4. The demographics and imaging information related to this cohort, which will be referred to as COVIDSet1 (CS1), are given in Table 5.1.

For CS1, to demonstrate that the medical resources needed to adequately treat patients who needed steroids and those who did not were notably different, Kaplan-Meier survival analysis was performed with time of hospitalization exchanged for time of survival (Figure 5.1) [132]. Particularly, patients who demonstrated a higher pneumonia severity index (PSI) grade experienced much longer hospitalization times than those with a lower PSI [133, 134]. This demonstrated the need for appropriate, consistent management and treatment of patients who progressed to severe disease stages, especially during peak resurgences of COVID-19 when a heavy burden is placed on medical systems to replenish and maintain resources [135, 136].

Table 5.1: COVIDSet1 Database Information

	Pre-treatment Analysis	During Treatment Analysis
Number of Cases	41 Scans	221 Scans (41 Cases)
Number of CT Scans Acquired (N)	NA	3 Scans (3 Cases) 4 Scans (7 Cases) 5 Scans (10 Cases) 6 Scans (14 Cases) 7 Scans (6 Cases) 8 Scans (1 Case)
Average Number of Timepoints	NA	Mean (5.39) SD (1.21)
Dates of Acquisition	Feb 01, 2020 - March 30, 2020	
Sex at Birth	Male (19) Female (22)	
Age	Mean(63.8) SD (11.5) Range(40-87)	
Scanner Manufacturer	GE Medical Systems	
kVp	120	
Pitch	Range (0.9844 - 1.750)	
Slice Thickness	0.625 mm, (211 Scans) 5 mm (10 scans)	

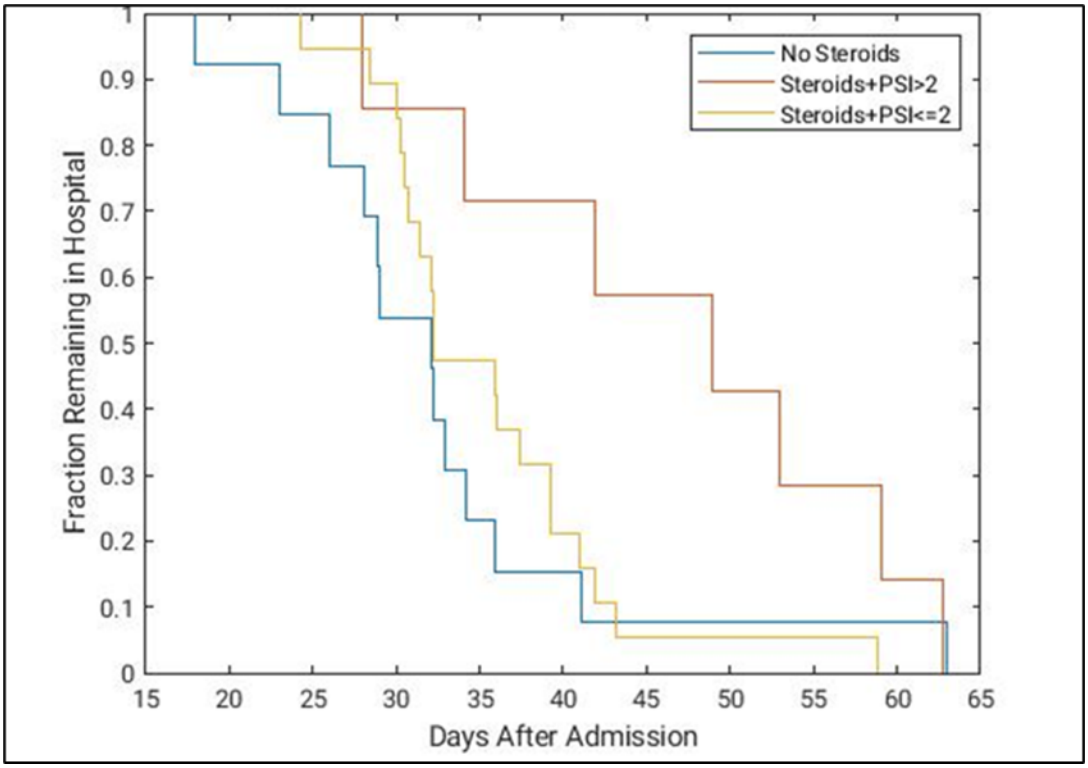

Figure 5.1: Kaplan-Meier survival analysis assessing the duration of hospitalization with changing treatment and initial PSI score. In general, patients who received steroid treatments were hospitalized for longer periods of time, with particularly long stays for patients with more severe initial symptoms. This is expected, as more severe cases require increased treatment and recovery time.

The second set of data, COVIDSet1 (CS2), was acquired over a much longer time period, November 2019 – December 2020, resulting in a much larger, more diverse data set. In contrast to the prior data, this cohort contained patients who did not progress to severe disease, had a larger range of outcomes (including patient death), and included additional comorbidity information that could be incorporated into machine learning models. The characteristics of this second cohort, COVIDSet2 (CS2), are described in Table 5.2.

5.2 Preliminary Study: Cascaded Transfer Learning with COVIDSet1

Early in the pandemic, administration of corticosteroids, particularly dexamethasone and methylprednisolone, served as the primary treatment option for patients who progressed to severe COVID-19 infection due to their impact in treating inflammatory symptoms in other respiratory infections [126, 127, 137]. However, patient reaction to steroid administration is variable, depending on many factors including patient age, smoking history, and other comorbidities. For COVID-19 treatment, thoracic imaging through chest CT may be used to inform treatment regimen, which is especially critical for COVID-19 due to the significant burden that this disease can place on medical resources.

As discussed in Section 2.3, transfer learning is a machine learning technique that allows for complex, hierarchical features to be extracted from imaging data, even in cases of limited data, by applying a pre-trained model to a new domain at the expense of utilizing potentially sub-optimal model parameters for the new domain task [4, 13]. One common strategy to overcome this is a process called fine tuning in which some subset of model parameters may be frozen (e.g., not trainable/adjustable) and the model is trained with a small learning rate to optimize the non-frozen parameters to the new domain task. Our first study utilized such a strategy, with a novel cascaded transfer learning technique for prognostic and temporal evaluations of CT scans obtained from COVID-19 patients in CS1.

Table 5.2: COVIDSet2 Database Information

Number of Cases	864 Cases 1842 Total Scans
Average Number of Timepoints	Mean (1.92) SD (1.19)
Dates of Acquisition	Nov 14, 2019 - Dec 4, 2020
Sex at Birth	Male (392) Female (472)
Age	Mean (53.5) SD (16.7) Range (8-90+)
Age Ranges	<30 (58 Cases) 30-65 (567 Cases) >65 (239)
Scanner Manufacturer	GE Medical Systems (929 Scans) Siemens (599 Scans) Philips (83) United (161) FMI (148)
Comorbidities	HBP (174 Cases) Liver Disease (87 Cases) Renal Disease (54 Cases) Neoplastic Disease (32 Cases) Cerebrovascular Disease (29 Cases) Congestive Heart Failure (13 Cases) COPD (7 Cases)
Patient Outcomes	Cured (261 Cases) Improving (526 Cases) Died (45 Cases) Other/Unknown (32 Cases)

The VGG19 architecture, a technique from the ImageNet competition, is commonly used in transfer learning through pre-training on a collection of millions of natural images. In Chapter 4, fine tuning of the ImageNet-trained VGG19 network had been conducted in the task of emphysema detection. In this COVID-19 study, CT slices were input to the fine-tuned VGG19 network and features were extracted for additional transfer learning using a technique similar to that described by Antropova et. al. in which information is taken from the max-pooling layers of the architecture (Figure 5.2a) [4]. These extracted features were then averaged in the axial scanning direction (e.g., individual slice features combined to form a CT-scan-level representation) and principal component analysis was performed on the extracted COVID-19 scan features for dimensional reduction to obtain the most dominant features corresponding to COVID-19. Final classification was then conducted on these dominant features using a support vector machine (SVM).

Due to the limited size of this CS1, a leave-one-out-by-case scheme was used to train the SVM for the classification between cases that required steroid administration and those that did not [138]. To attain a prognostic evaluation of COVID-19 patients, only the initial CT scan obtained for each patient was evaluated, at which point the patients presented with varying degrees of disease severity. The leave-one-out evaluation approach used 40 cases for SVM training and 1 case for testing; this was repeated 41 times so that each case belonged to the testing set exactly once. Over the 41 iterations, the SVM produced an output "prediction score", related to the likelihood of requiring steroid treatment, for each case. The prediction score yielded an estimate of the likelihood that a patient would require steroids for treatment based on their CT scan (higher prediction score indicates higher likelihood of recommendation for steroid treatment). The classification performance was evaluated using ROC analysis on the prediction scores by comparison with the actual treatment as had been clinically determined by an expert intensivist. The AUC served as the figure of merit in this

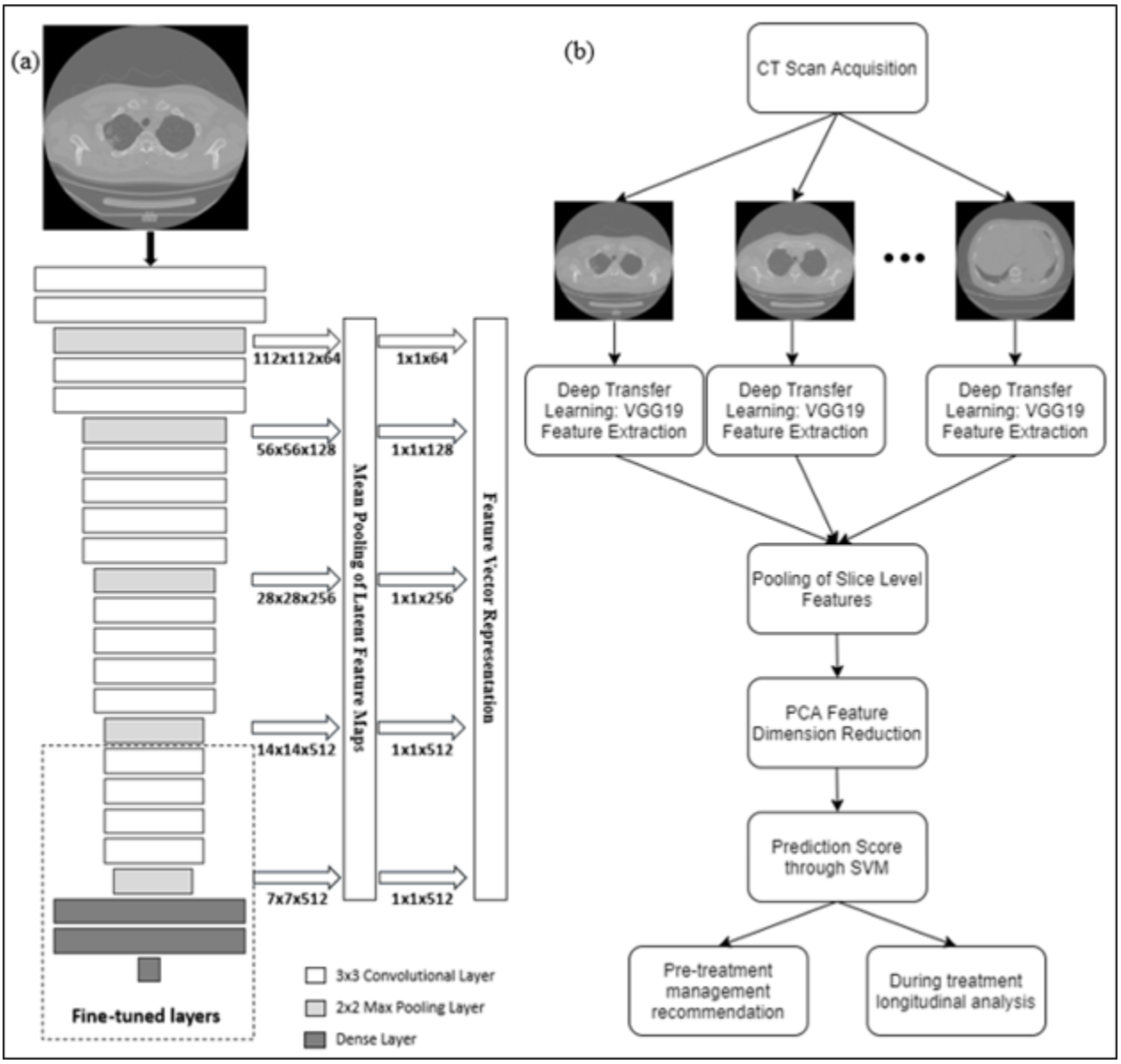

Figure 5.2: (a) Schematic of the pretrained VGG19 network feature extraction approach operating on a 2D CT section. Max pooling layer features with the given dimensions were averaged and concatenated to produce a representative feature vector for each slice. (b) Full cascaded transfer learning workflow for pre-treatment assessment and during-treatment monitoring analysis. The feature extraction scheme displayed in (a) is utilized at the "Deep Transfer Learning: VGG19 Feature Extraction" stage of (b).

analysis. The full workflow of the cascaded transfer learning technique is shown in Figure 5.2.

In addition to predicting which patients should require steroids, it is also important to monitor disease progression throughout hospitalization. Thus, the cascaded transfer learning technique was also applied to the CT scans obtained at all timepoints; the only difference between this during-treatment assessment technique and the pre-treatment assessment technique is that the linear SVM classifier for during-treatment assessment was trained using features from all longitudinally-acquired CT scans, not only the initial scan of each case.

After each of the CT scans within a case underwent deep transfer learning feature extraction and subequent PCA feature reduction, the SVM was trained again using the leave-one-out-by-case paradigm based on receiving steroids or not with incorporation of all timepoints in both training and testing. Thus, each patient received a set of model prediction scores, one prediction score for each timepoint's CT scan. The time of steroid administration was eliminated as a confounding factor by adjusting the class label for a given scan based on whether or not steroids were utilized at any point after that scan's acquisition. For example, a case with 5 timepoints could have a timepoint class label of 1 at timepoints 1, 2, and 3 and class label of 0 at timepoint 4 and 5, indicating no steroids were administered after the 4th scan acquisition; this can be represented as a set 1,1,1,0,0. The SVM prediction score at a mid-treatment timepoint can then be interpreted as a prediction of whether the patient will undergo steroid treatment at any point after that mid-treatment scan. Note that during the leave-one-out process, all CT scans of a given case were held out from training when it was used for testing.

The assessment of temporal changes throughout hospitalization was performed through least squares fitting. All patients within the study cohort began with moderate severity and advanced to a more serious condition, followed by recovery and subsequent hospital discharge. This was observed for both cases who were treated with steroids and those who

were not. Thus, the least squares technique was used to fit second-order polynomials which, to some degree, match the expected progression for both groups.

5.2.1 Results from ROC Analysis and Temporal Analysis

In the predictive analysis of the initial (pre-treatment) CT scans, the cascaded transfer learning technique produced an AUC of $0.85+/-0.10$ based on proper binormal ROC analysis in the task of distinguishing between cases that were recommended for steroid administration and those that were not, demonstrating a statistically significant improvement in comparison to a random chance AUC of 0.5 ($p = 0.002$) (Figure 5.3a). By analyzing the distribution of the deep learning scores based on the true steroid administration (Figure 5.3b-c), there were 2 outliers within the distribution of cases that received steroid treatments. These two outliers belonged to patients with low PSI scores and were young compared to the mean population age (ages 41 and 48).

Preliminary longitudinal analysis was completed through least squares fitting of the raw data (Figure 5.4). Due to the variable initial disease state, rate of progression, and treatment schedules, there was substantial variation across patients, thus the wide coverage of the shaded regions denoting a one standard deviation range above and below the fit line. Based on ROC analysis in the task of identifying patients who required steroid treatments or not, and the longitudinal trends obtained through least squares fitting, the cascaded transfer learning approach showed strong potential for clinical patient management through informing treatment decisions and monitoring patient progression.

While preliminary, this technique demonstrated potential to estimate a likelihood that a patient will progress to a disease stage that is severe enough to necessitate steroid administration during their course of treatment. This holds potential value for allowing hospitals to obtain sufficient medical resources for adequate patient care, including maintenance of steroid supplies, utilization of life-saving equipment such as ventilation, extracorporeal membrane

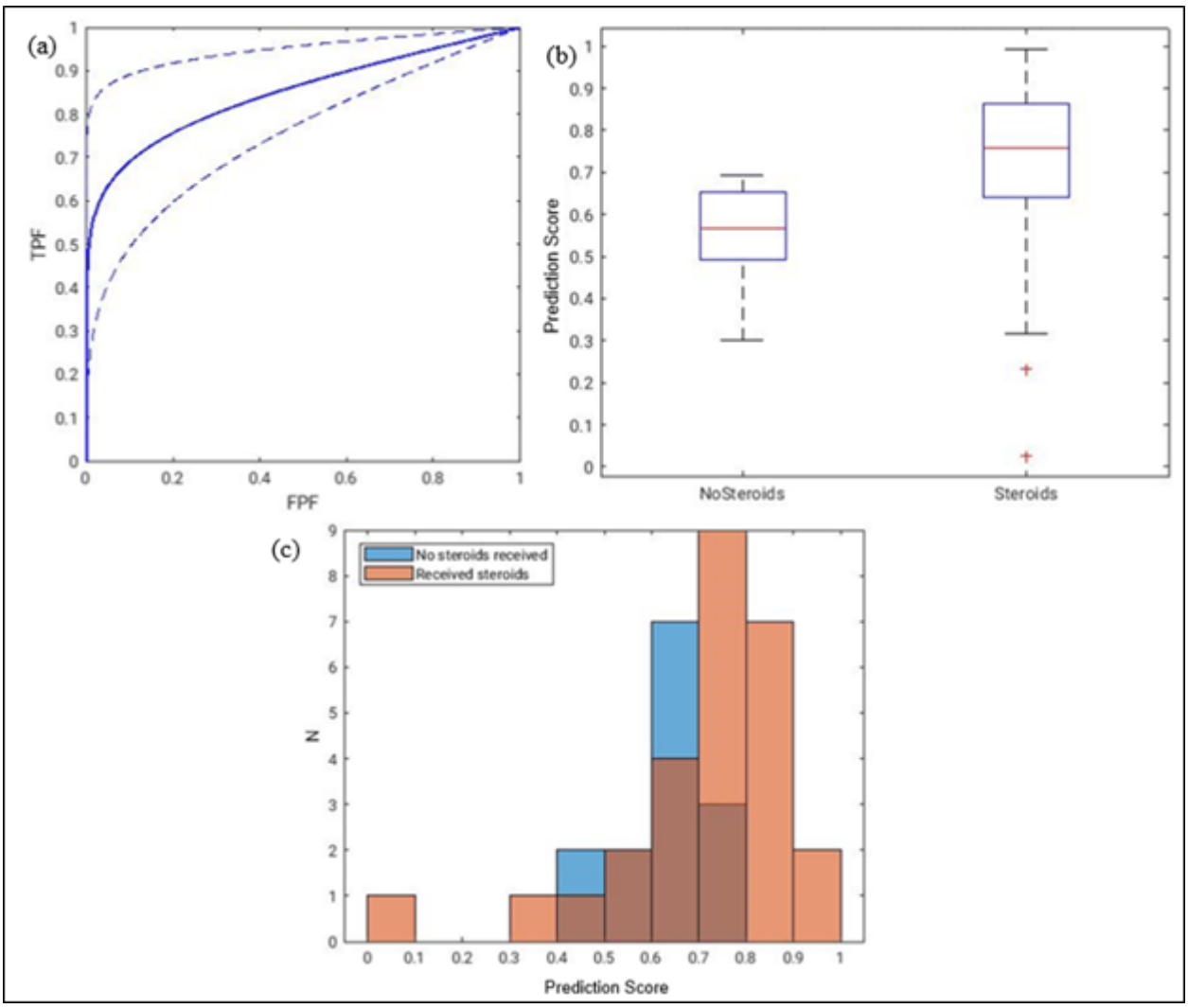

Figure 5.3: (a) The ROC curve demonstrating the classification ability of the cascade transfer learning method for estimating the likelihood that a COVID-19 patient would be recommended for steroid treatment or not. AUC $= 0.85 + / - 0.10$ with the accompanying 95% TPF confidence interval. (b) Distribution of deep learning scores of those patients who received steroids and those who did not. Note, this was obtained only based on the initial CT scan. Based on this plot, the method suggests steroid administration more frequently than the experienced intensivist (using a cutoff of 0.5). The red lines denote the median scores, the blue boxes include 50% of scores, while the black whiskers include all scores within 2σ of the mean. (c) Further demonstration of the separation/overlap of the deep learning score between the two classes.

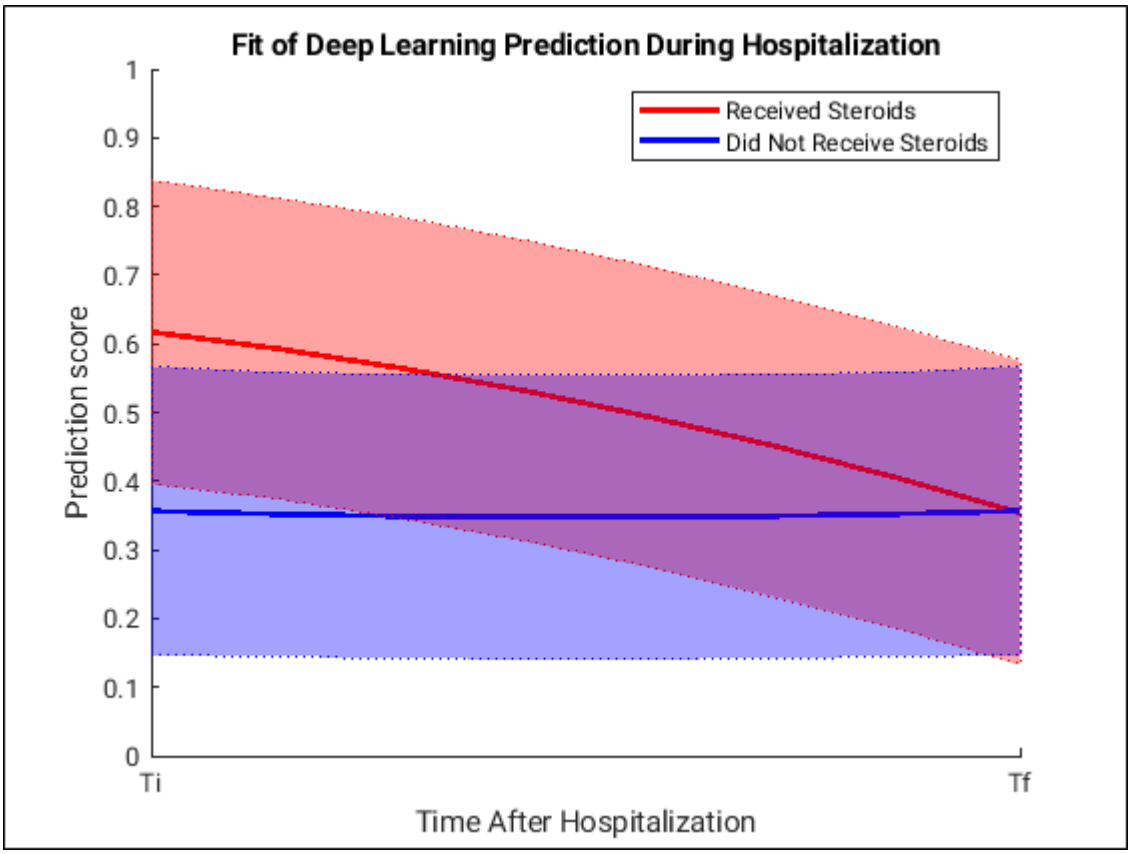

Figure 5.4: The SVM-output prediction score assessed temporally through least squares fits. The x-axis indicates the full duration of hospitalization, with Ti referring to the time of initial CT acquisition and Tf referring to the time of final CT acquisition, which generally occurred shortly before discharge. The shaded regions denote one standard deviation above and below the fit line. Intuitively, this figure follows the example training case discussed in Section 2.3 which had early timepoints after which steroids were utilized and late acquisitions after which no steroids were administered.

oxygenation, and planning for hospital bed occupancy. These are challenges that stressed the medical system during the COVID-19 pandemic, thus this predictive tool could be useful in future resurgences of COVID-19 or other emerging respiratory infectious diseases. Further, this deep transfer learning technique might not only benefit allocation of resources, it could also potentially improve patient management through treatment guidance as indicated by the temporal results matching the treatment decision of the experienced intensivist who provided the reference standard in this study.

Use of deep transfer learning for monitoring treatments may be useful for estimating the amount of resources that will be required throughout patient treatment as demonstrated in our preliminary longitudinal analysis shown in Figure 4, including hospital beds, ventilators, and medication. The SVM prediction score for this analysis may be interpreted as an expectation of whether that patient will require steroids at any point after input CT acquisition. Consequently, if a patient is already experiencing corticosteroid injections for treatment and receives a high prediction score, this suggests that they should continue steroid treatments for some period.

Importantly, the fit lines can not be used as a direct indicator of disease severity obtained at any individual timepoint during this study; if this were the case, then the fit lines would both demonstrate, on average, a concave mirroring effect showing that the prediction score increases as time progresses, and then decreases as treatment takes effect until a very low level is reached as the disease subsides. This expectation of concave mirroring prediction score is due to the disease progression manifested by radiological findings of COVID-19 in CT scans; with increasing severity, the CT findings exhibit GGOs, more central infiltration, and consolidation, then a return to primarily peripheral GGO as the patient recovers. This mirroring effect is not observed for either classification. The failure to demonstrate a mirroring effect for fit line from steroid administration can be explained by the class label used for model training, which was a decision to treat with steroids or not at some point after scan

acquisition. While this is inherently related to disease severity, this treatment decision was not solely based on imaging findings; it also considered clinical symptoms and other factors (e.g., age). Thus, the fit line does not directly translate to a temporal assessment of severity during this study.

Instead, the trends should be interpreted as a recommendation for steroid administration. On average, the cases that required steroids demonstrated a larger prediction score for scans acquired upon hospitalization. However, as time after hospitalization increased, the two curves converged, indicating that the recommendation for steroid administration grew weaker over time for those who already received steroids, and demonstrated nearly complete overlap at the termination of hospitalization. This matched expected results because the clinical outcomes of all patients in this cohort were the same, i.e., recovery and subsequent discharge, thus all patients should demonstrate a similar recommendation for steroids upon discharge. This validated the prediction score as a potentially useful clinical measurement.

Clinically, the longitudinal aspect of this study can be used by physicians as a comparison to guide treatment decisions and, in a way, assess treatment response (e.g., is the model recommendation for steroid administration getting stronger or weaker?). Further, consider a patient that has been administered corticosteroids that now produces a CT scan with a low prediction score (e.g., 0.3). According to the temporal fits in Figure 4, this suggests that the patient is likely nearing the end of their hospitalization period and that cessation of steroid administration may be suitable. Alternatively, it is possible that some patients will not follow a progression of prediction scores similar to the fit lines; in this case, the temporal assessments may not be applicable and the clinician should be more reliant on other data (e.g., clinical symptom severity) to determine steroid treatment termination. Thus, the cascaded transfer learning approach in this study demonstrated potential in guiding treatment decisions, monitoring patient progression, and managing medical resources.

5.3 Further Study: Deep Learning with COVIDSet2

As discussed in Section 1.2, it is common for machine learning algorithms to be biased to the training set distribution. Thus, we directly applied the model developed on CS1 to CS2 to determine if it would successfully generalize to a larger population.

Despite the strong performance achieved in the preliminary study, the developed prognostic model failed to reach the same performance on the larger, more diverse CS2. In direct application for the same pre-treatment prediction task, the AUC was reduced from 0.85 +/- 0.10 to 0.68 +/- 0.04 (p=0.02); while this AUC did maintain a statistically significant difference from guessing AUC of 0.5 (p=0.0001) it likely is no longer a clinically relevant performance. There are several potential contributing factors, including potential treatment decision differences between the acquisition periods in CS1 and CS2, the variability between individuals making treatment decisions (e.g., only one clinician provided treatments in CS1 while CS2 had many different leading clinicians making those decisions), the more diverse imaging acquisition characteristics including scans acquired from devices from different scanner manufacturers (Table 5.2). Further, there could be an algorithmic shortcoming; cross validation tends to overestimate performance compared to a truly independent train, validation, and test strategy; thus, the independent evaluation conveyed a more realistic clinical performance. Regardless of the root cause, the model no longer achieved promising results.

In an attempt to develop a more generalizable model with comparable performance to the initial study, three potential improvements were investigated: 1) evaluation of improved feature extraction models, 2) investigation of novel slice-to-scan pooling methods (including the attention-based pooling presented in Section 5.2), and 3) incorporation of additional, non-imaging information in the form of comorbidity data. Thus, we performed three further studies to determine the best feature extraction, pooling, and fusion methods for the steroid prediction task.

In Chapter 4 all transfer learning models utilized for CT slice feature extraction were initialized with pre-trained weights from the ImageNet classification challenge which, as discussed in Section 2.3, is potentially suboptimal for application in the new task domain. In the prior study, the cascaded transfer learning approach aimed to bridge the domain gap by tuning the model for the LDCT emphysema classification task; while this was an improvement that allowed transfer to the medical imaging domain, the transferred model prioritized information that was specific to the emphysema imaging task and may have been biased for other imaging characteristics of the LDCT dataset, including presentation of emphysema, lower acquisition dose, and single scanner manufacturer, and thus extracted suboptimal features for the COVID-19 classification task. Recently, a set of medical imaging transfer learning models, RadImageNet, were published to facilitate application of transfer learning technology in medical imaging [139]. These models can replace standard ImageNet-trained models (e.g., ResNet50, DenseNet121) in machine learning workflows and are trained directly for medical imaging classification tasks, potentially improving feature relevance for classification. As opposed to the original model, the RadImageNet dataset consists of images from several different medical imaging modalities, including radiography, MRI, and CT, and thus the pre-trained models aim to extract imaging features that are generally informative and can utilize newly trained classifiers for specific tasks. In the original publication, the RadImageNet models outperformed ImageNet models for several transfer learning tasks in the medical imaging domain for direct classification tasks. Thus, in this revised study, we evaluated if RadImageNet models would also outperform ImageNet models for a more complex machine learning paradigm, i.e., MIL, in the task of COVID-19 patient prediction of steroid administration.

While the ImageNet to RadImageNet transition impacted the feature extraction stage of MIL, we also explored alternatives to the CT slice pooling stage. In Chapter 5, we utilized attention-based pooling to incorporate a learned, optimized method for aggregating slice

information, and here we use attention-based pooling again for the COVID-19 prediction task [5]. Additionally, we investigated a more complex, novel approach to slice aggregation through a transformer-based self-attention module.

5.3.1 Transformer Architectures

In recent years, transformer architectures have dominated the natural language processing deep learning space and have begun to experience increased application in computer vision tasks [140, 141]. The fundamental unit of the transformer architecture is self-attention, or an attention mechanism which evaluates how different parts of a sequence interact and influence each other. The application of such a technique in natural language processing is intuitive; a word may impact the meaning/interpretation of other words in the sequence. We apply this same technique to multiple instance learning pooling by treating the extracted CT slice features similarly to word embeddings and allow the attention module to learn the optimal method for the slice features to interact and determine which slices are/are not important for the classification task [117]. This is achieved by mapping each CT slice to learned query (Q), key (K), and value (V) embeddings which, in brief, represent how a slice should influence other slices (Q), how a slice should be influenced by the Q embeddings (K), and how the information from the original embedding should be carried to the next layer (V). This process is visualized in Figures 5.5 and 5.6, including how several self-attention layers can work in series to form the deep network architecture utilized in this study.

5.3.2 Incorporation of Human-Engineered Features and Clinical Presentation of COVID-19

The final adjustment from the prior study came in the form of adding non-deep learning information, i.e., human-engineered radiomic features and clinical comorbidity data. Two major shortcomings of the original model were the lack of non-imaging clinical information

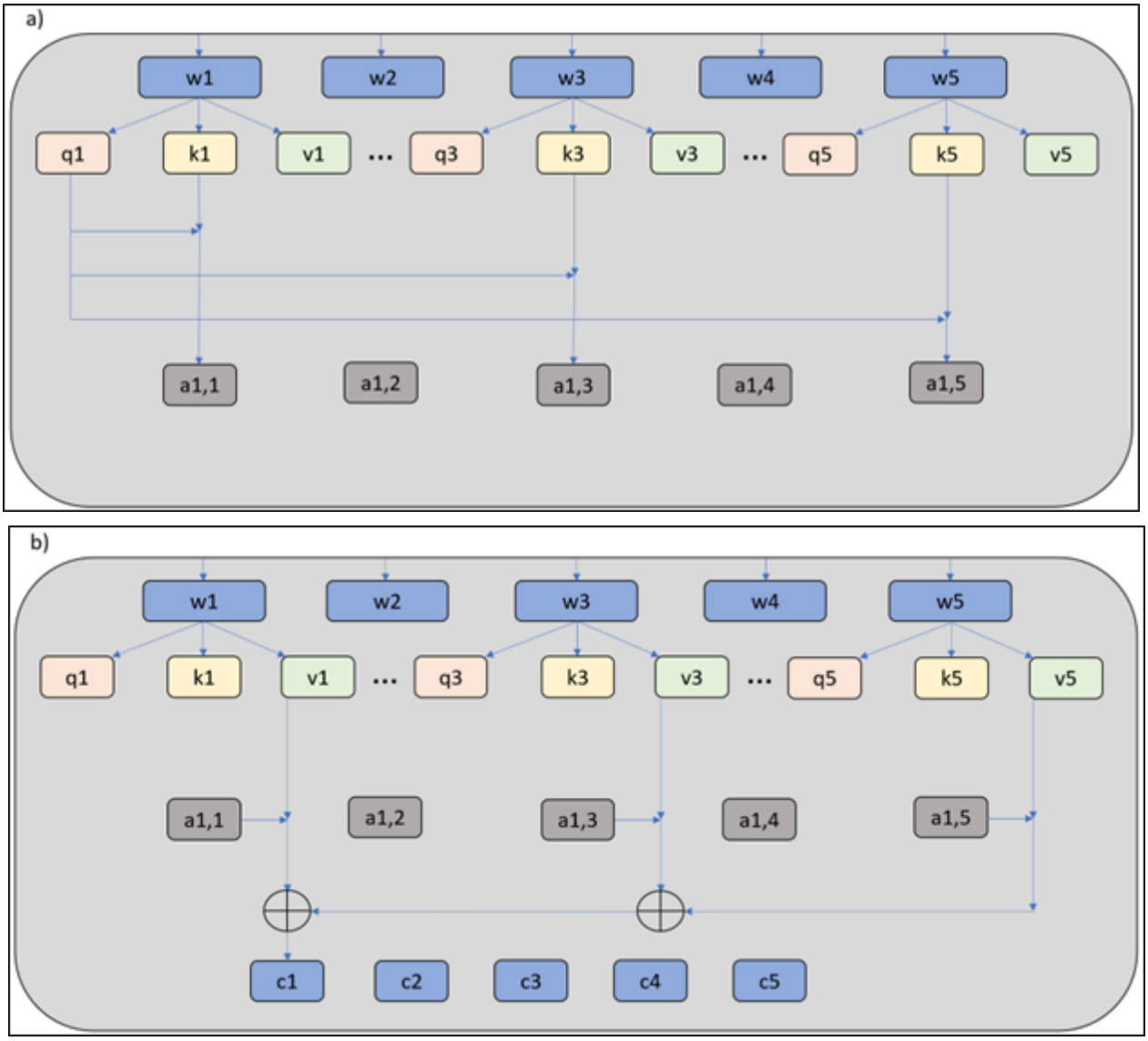

Figure 5.5: Depiction of the self-attention module utilized in transformers a) Demonstrates the generation of attention weights for slice 1 via mapping each input representation w to Q and K representations, then interacting the relevant Q representation (in this case, q1) with the K representations of other slice representations. b) Shows the aggregation of attention weights with each V representation to form the output representation of slice 1, c1.

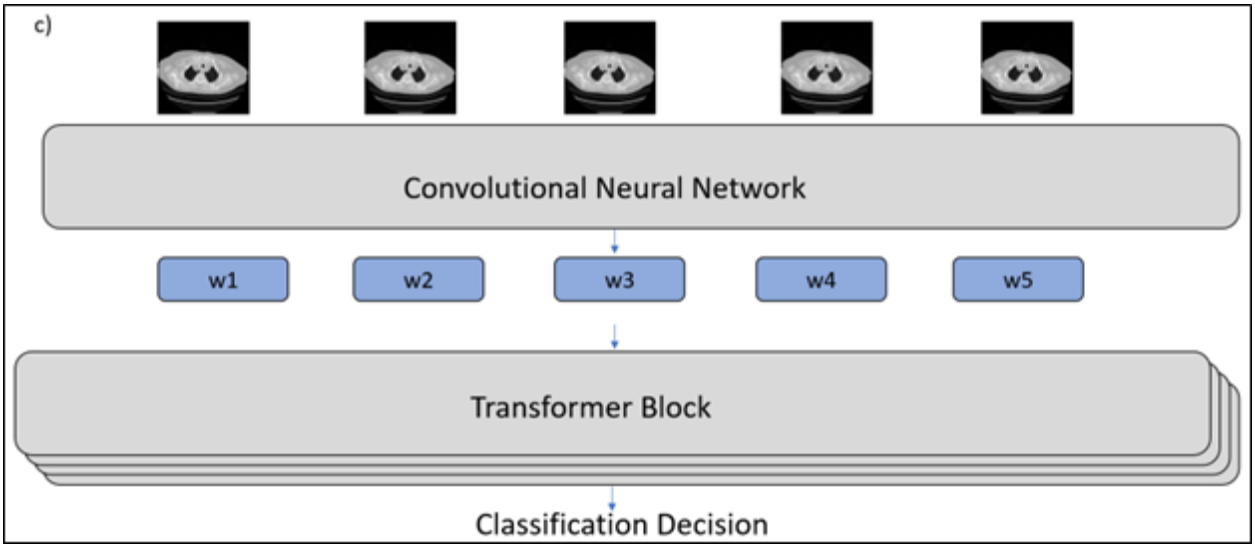

Figure 5.6: Depiction of how transformers may be utilized for CT scan evaluation. By first embedding the images to a feature representation via a convolutional neural network, self-attention modules can be utilized in parallel and in series to evaluate the image features and form a classification decision.

(which would generally be considered in a treatment decision) and the lack of model interpretability. We account for both of these by developing additional models for prediction and feature fusion with the MIL approach and evaluate how these additional models impact the steroid prediction performance.

First, we automatically segmented the lungs and presentation of COVID-19 within the lungs on individual CT slices utilizing a novel dual-headed U-Net architecture. Intensity-based human-engineered features were extracted from the segmentations, e.g., mean pixel value and ratio of COVID-19 involved tissue to healthy lung tissue. In past studies, deep network features have been shown to outperform radiomics features for several medical imaging tasks at the expense of interpretability; here, we only include features that were readily interpretable by a clinical reader and will add interpretability to the greater MIL model. These features will be referred to as segmentation-based radiomic (SBR) features, with the full extraction workflow given in Figure 5.7.

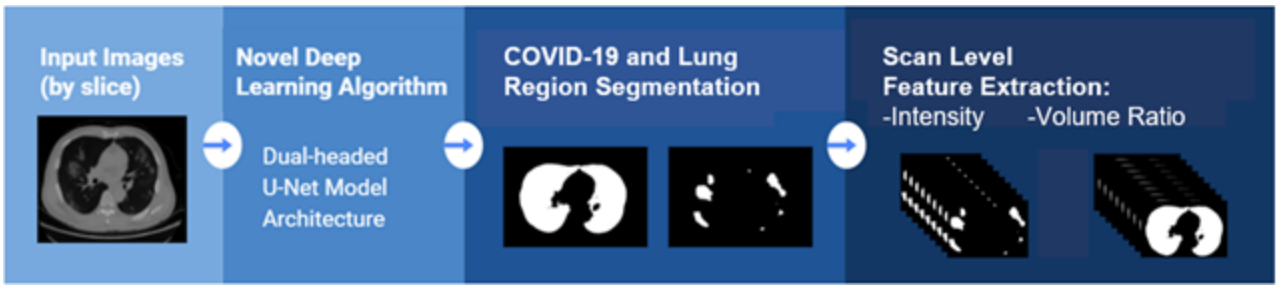

Figure 5.7: Workflow for the extraction of SBR features from CT scans. We first segment the lung and COVID-19 infection within individual CT slices using a novel dual-headed U-Net model, then extract intuitive features including intensity-based and volume ratio features. The full list of features is provided in Table 5.3.

As previously mentioned, comorbidities play a key role in identifying individuals at high risk for severe COVID-19 infection [142, 143]. Thus, the second additional source of non-deep learning information was tabular, binary comorbidity data indicating diagnoses of high blood pressure (HBP), liver disease, renal disease, neoplastic disease, cerebrovascular disease, congestive heart failure, and chronic obstructive pulmonary disease (COPD). These features will be referred to as tabular comorbidity (TC) features. The full list of SBR and TC features is given in Table 5.3.

The MIL, SBR, and TC features were aggregated in two approaches, prediction fusion and feature fusion. Prediction fusion trained individual classifiers for each feature type to individually predict the COVID-19 steroid administration task. Notably, different classifier approaches are appropriate for each feature type, thus the trained classifiers were 1) a random forest for the TC features, 2) a SVM for the SBR features, and 3) an artificial neural network for the MIL features. The individual predictions were then averaged to determine the scan prediction. Additionally, a feature fusion scheme in which all features were concatenated into a single, scan level representation and classified utilizing an artificial neural network was compared. This workflow is depicted in Figure 5.8.

Table 5.3: List of Incorporated SBR and TC Features for COVID-19 Steroid Prediction Task

Feature Name	Target and Side of Body	Feature Category
Age	N/A	TC
COPD	N/A	TC
HBP	N/A	TC
Congestive Heart Failure	N/A	TC
Cerebrovascular Disease	N/A	TC
Renal Disease	N/A	TC
Liver Disease	N/A	TC
Neoplastic Disease	N/A	TC
Pixel Mean	Lung/Left	SBR
Pixel Mean	COVID-19/Left	SBR
Pixel Mean	Lung/Right	SBR
Pixel Mean	COVID-19/Right	SBR
Pixel Mean	Lung/Both	SBR
Pixel Mean	COVID-19/Both	SBR
Pixel Standard Deviation	Lung/Left	SBR
Pixel Standard Deviation	COVID-19/Left	SBR
Pixel Standard Deviation	Lung/Right	SBR
Pixel Standard Deviation	COVID-19/Right	SBR
Pixel Standard Deviation	Lung/Both	SBR
Pixel Standard Deviation	COVID-19/Both	SBR
Pixel Minimum	Lung/Left	SBR
Pixel Minimum	COVID-19/Left	SBR

Table 5.3: Continued

Feature Name	Target/Side of Body	Feature Category
Pixel Minimum	Lung/Right	SBR
Pixel Minimum	COVID-19/Right	SBR
Pixel Minimum	Lung/Both	SBR
Pixel Minimum	COVID-19/Both	SBR
Pixel Maximum	Lung/Left	SBR
Pixel Maximum	COVID-19/Left	SBR
Pixel Maximum	Lung/Right	SBR
Pixel Maximum	COVID-19/Right	SBR
Pixel Maximum	Lung/Both	SBR
Pixel Maximum	COVID-19/Both	SBR
Pixel 25% Distribution	Lung/Left	SBR
Pixel 25% Distribution	COVID-19/Left	SBR
Pixel 25% Distribution	Lung/Right	SBR
Pixel 25% Distribution	COVID-19/Right	SBR
Pixel 25% Distribution	Lung/Both	SBR
Pixel 25% Distribution	COVID-19/Both	SBR
Pixel 75% Distribution	Lung/Left	SBR
Pixel 75% Distribution	COVID-19/Left	SBR
Pixel 75% Distribution	Lung/Right	SBR
Pixel 75% Distribution	COVID-19/Right	SBR
Pixel 75% Distribution	Lung/Both	SBR
Pixel 75% Distribution	COVID-19/Both	SBR
Ratio: diseased tissue volume to total lung volume	Left	SBR

Table 5.3: Continued

Feature Name	Target/Side of Body	Feature Category
Ratio: diseased tissue volume to total lung volume	Right	SBR
Ratio: diseased tissue volume to total lung volume	Both	SBR

5.3.3 Statistical Data Analyses

Similar to the prior study, the prediction task was evaluated through ROC analysis with the AUC as the performance metric, statistical comparison through the DeLong test, and the Bonferroni-Holm correction was applied to account for multiple comparisons [103, 144].

5.4 Results on COVIDSet2

In repeating the pre-treatment task of predicting which patients would require steroids in CS2, the results are shown displayed in Tables 5.4. The first study section to determine the utility of RadImageNet pre-training compared to ImageNet pre-training revealed detrimental effects introduced by the RadImageNet model substitution. The top-performing feature extractor was ResNet50 with ImageNet pretraining, statistically significantly outperforming the RadImageNet counterpart by ΔAUC of 0.12 ($p \ll 0.001$) while the DenseNet121 also demonstrated a statistically significant difference between the ImageNet and RadImageNet pretraining with $\Delta AUC = 0.04$ ($p \ll 0.001$). Because ResNet50 demonstrated the best performance, it was utilized as the feature extractor for the other ablation analyses.

Similarly, the incorporation of vision transformer modules for pooling failed to demonstrate an improvement compared to attention-based pooling (Table 5.6) with a performance

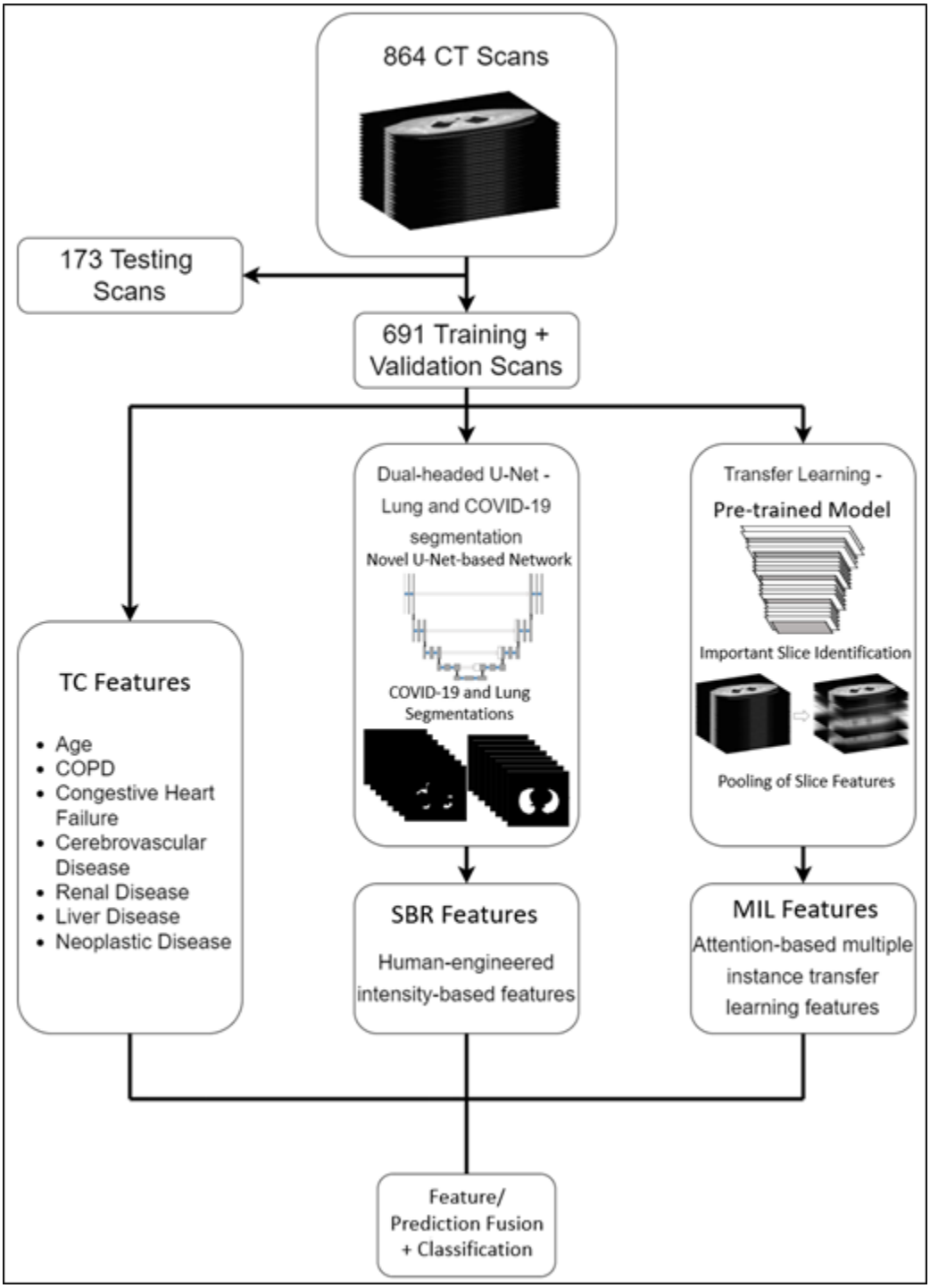

Figure 5.8: Full feature fusion workflow of the MIL, SBR, and TC feature pipeline for each of the 5 cross validation folds. The artificial neural network classifier can be replaced with individual classifiers for each feature type to visualize the prediction fusion approach.

Table 5.4: Comparing ImageNet and RadImageNet Feature Extraction

Feature Extraction Architecture	Pre-training Weights	AUC
ResNet50	ImageNet	0.76 +/- 0.08
ResNet50	RadImageNet	0.64 +/- 0.09
DenseNet121	ImageNet	0.68 +/- 0.12
DenseNet121	RadImageNet	0.64 +/- 0.09

Table 5.5: Statistical testing for ImageNet vs. RadImageNet (p-values); bold indicates statistical significance

Model/ Pre-training Weights	ResNet50/ ImageNet	ResNet50/ RadImageNet	DenseNet121/ ImageNet
ResNet50/RadImageNet	<<**0.001**	-	-
DenseNet121/ImageNet	<<**0.001**	<<**0.001**	-
DenseNet121/RadImageNet	<<**0.001**	0.10	<<**0.001**

difference of $\Delta AUC = 0.25$ ($p \ll 0.001$). Transformer pooling failed to show a statistical different from a random guessing AUC.

Table 5.6: Transformer pooling vs. attention-based pooling

Pooling Mechanism	AUC
Attention-based Pooling	0.76 +/- 0.08
Transformer Pooling	0.51 +/- 0.06

Finally, when incorporating the additional SBR and TC features through feature fusion and prediction fusion methods, the prediction task again failed to produce a statistically significant difference compared to the use of MIL features alone (Table 5.7). Individually, the MIL model outperformed the TC features by a statistically significant difference of $\Delta AUC = 0.21$ ($p \ll 0.001$). In Figures 5.9 and 5.10, we compare the model predictions for different selections of Table 5.7 to determine the influence of incorporating different feature types.

Table 5.7: Incorporating additional feature types for steroid treatment classification prediction

Included Feature Types			Model Number	Fusion Method	AUC
MIL: Deep Features	TC: Comorbidity Features	SBR: Radiomics Features			
X			1	None	0.76 +/- 0.08
	X		2	None	0.55 +/- 0.04
		X	3	None	0.74 +/- 0.06
X	X		4	Feature	0.77 +/- 0.08
X		X	5	Feature	0.76 +/- 0.07
X	X	X	6	Feature	0.77 +/- 0.07
X	X		7	Prediction	0.78 +/- 0.08
X	X	X	8	Prediction	0.74+/- 0.04

Table 5.8: Statistical testing for incorporation of additional feature; bold indicates statistical significance; Model numbers listed in Table 5.6

Model Number	1	2	3	4	5	6	7
2	<<**0.001**	-	-	-	-	-	-
3	0.058	<<**0.001**	-	-	-	-	-
4	0.69	<<**0.001**	0.048	-	-	-	-
5	0.87	<<**0.001**	0.60	0.69	-	-	-
6	0.14	<<**0.001**	0.21	0.078	0.083	-	-
7	0.0056	<<**0.001**	<<**0.001**	0.15	0.95	**0.0016**	-
8	0.023	<<**0.001**	0.75	0.022	0.032	0.22	<<**0.001**

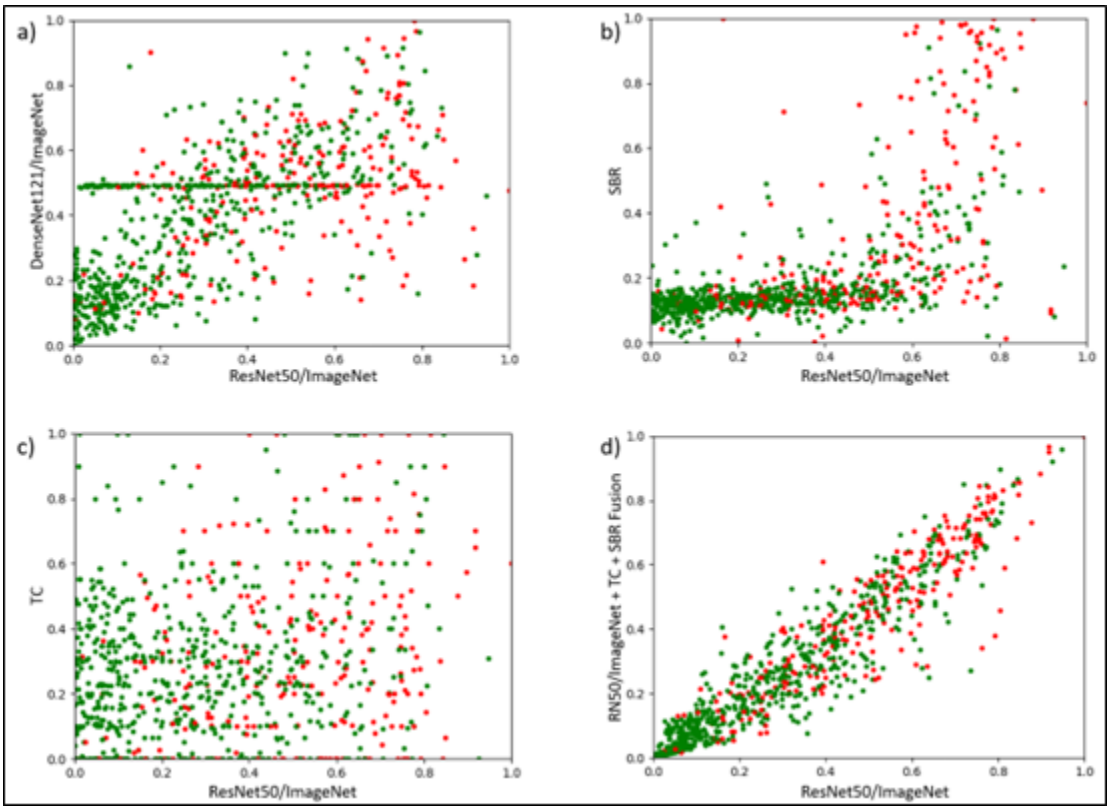

Figure 5.9: Comparison of predictions in the task of predicting COVID-19 steroid administration based upon initial patient CT scan and Bland-Altman plots from different models in ablation studies with patients who received steroids (red) and those who did not (green). (a) Compares predictions between the two ResNet50 and DenseNet121 models trained with ImageNet pre-training and attention-based MIL pooling (Table 5.4). Note that in one of the 5 cross-validation folds, the DenseNet121 model failed to converge to a useful set of parameters and the model produced a constant output regardless of input. This is observed with the unusual line of points through the center of the plot. (b) Compares the MIL model with SBR model. While both achieved similar AUCs, the SBR model seems to be recall cases at a much lower rate. (c) Comparing the same MIL model with the TC model. The poor performance of the TC features is prevalent here with relatively little structure visualized here. (d) Comparing the MIL model to the same model with feature fusion. The fused model tends to slightly underpredict cases compared to the model using MIL features alone; this may potentially be credited to the inclusion of the SBR features in the fused model.

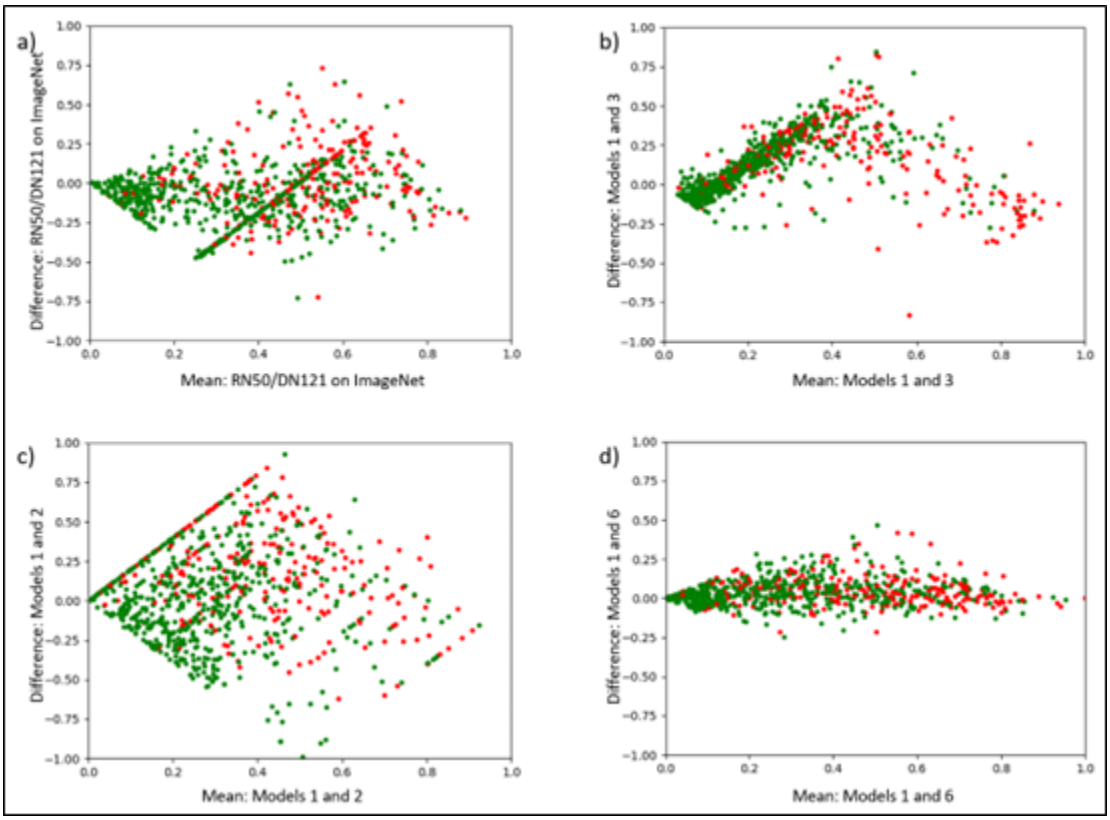

Figure 5.10: Comparison of predictions in the task of predicting COVID-19 steroid administration based upon initial patient CT scan and Bland-Altman plots from different models in ablation studies with patients who received steroids (red) and those who did not (green). (a) Compares predictions between the two ResNet50 and DenseNet121 models trained with ImageNet pre-training and attention-based MIL pooling (Table 5.4). Note that in one of the 5 cross-validation folds, the DenseNet121 model failed to converge to a useful set of parameters and the model produced a constant output regardless of input. This is observed with the unusual line of points through the center of the plot. (b) Compares the MIL model with SBR model. While both achieved similar AUCs, the SBR model seems to be recall cases at a much lower rate. (c) Comparing the same MIL model with the TC model. The poor performance of the TC features is prevalent here with relatively little structure visualized here. (d) Comparing the MIL model to the same model with feature fusion. The fused model tends to slightly underpredict cases compared to the model using MIL features alone; this may potentially be credited to the inclusion of the SBR features in the fused model.

5.5 Discussion and Conclusions

In the analysis of CS2, the best-performing model utilized a ResNet50 architecture for feature extraction, attention-based MIL feature pooling, and utilized the SBR features in a prediction fusion scheme achieving an AUC of 0.78 +/- 0.08. The primary gains in model performance seem to stem primarily from the selection of feature extraction architecture and the pooling mechanism while the additional non-MIL feature incorporation did not impact AUC regardless of feature fusion method.

Surprisingly, the use of RadImageNet as a feature extractor did not improve performance compared to ImageNet pre-trained models in either the ResNet50 or DenseNet121 models despite reduced shift between the original task domain (e.g., 2D medical image classification for several modalities) and the new application domain. We hypothesize three potential causes. First, the RadImageNet database consists of only 5 million images spread across several imaging modalities (radiographs, PET, MRI, CT, ultrasound) compared to the larger ImageNet database of 14 million images; it is feasible that the reduced number of training images affected the generalizability of the features. Second, it is likely that differences in image preprocessing (e.g., windowing, normalization, etc.) existed between the original RadImageNet data and the application domain, particularly considering it would be impossible to process the non-CT images in RadImageNet identically to this task. Finally, the RadImageNet features may be unsuited for extension to the non-standard MIL scheme utilized in this task. Typically, MIL schemes train the model end-to-end from scratch, including the feature extractor in the training process, and while we have shown that strong performance can be achieved in Chapter 4 with fixed ImageNet model parameters, the same generalizability of RadImageNet parameters is yet to be demonstrated.

In the second stage ablation study, the transformer module was unable to match the performance of the attention-based approach. As a relatively new technology still being explored for the natural and medical imaging spaces there have been some examples of successful use

of transformers in MIL, but there are several potential causes for the poor performance for transformer pooling. Foremost, transformers are much more heavily parameterized than the attention-based pooling method; thus, data limitations may have played a role in the performance loss. Further, the transformers utilized in this study consisted of many more parameters than typical vision transformers per layer; a layer is typically limited in either the number of tokens (i.e., CT slices) or the size of those tokens (i.e., the number of features extracted per image in the prior stage). For example, the original vision transformers only utilized 9 input tokens, whereas our study utilized >150 across experiments with comparable token size. Transformer variants that attempt to mitigate effects caused by such a phenomenon have been explored in the NLP community (e.g., Nystromformers), which we will explore for this task in future work [145].

The utilization of additional feature information failed to demonstrate a statistical difference in model performance according to the DeLong Test. Individually, the TC features did not demonstrate strong predictive power, thus limiting their potential impact on the fused model prediction performance. However, the SBR features were able to achieve strong individual performance that failed to demonstrate a statistically significant difference from the MIL models ($p = 0.58$). Evaluating Figures 5.9 and 5.10, we see that there is strong agreement between the ImageNet-based feature extraction techniques, but a failure to converge in one of the cross-validation folds drastically reduced the DenseNet121 performance and resulted in the consistent line of points around y=0.5 stretching from x=0 to x=0.7. This demonstrates a potential weakness of the MIL approach compared to the SBR and TC approaches; with a lack of interpretability, it is difficult to determine why the model failed to converge for only one of the 5 training scenarios.

In Figures 5.9 and 5.10 (b) and (c), the SBR features demonstrated a potential improvement compared to the MIL model while TC features seem to contribute only minimally to model performance. If a clinician desired a more conservative approach to steroid adminis-

tration, the SBR model may outperform the MIL model as very few false positives would be called at a decision threshold of 0.5; this would come at the expense of false negatives as well (as depicted by the many red points below the 0.5 threshold).

Finally, in Figures 5.9 and 5.10 (d), we compared the MIL model with the feature fusion model consisting of all three feature types. Notably, the fused model predictions were slightly reduced compared to the MIL features alone, but much closer than the difference observed in 5.9 and 5.10 (b). This difference may be attributed to the inclusion of the SBR features in the fused model, but the strong agreement for many of the cases suggests that the MIL features were the dominant feature in the fused model prediction. In the feature fusion model, the number of features contributed by TC (8) and SBR (39) were fewer than the number of MIL features (64) and were not trained end-to-end for optimization; this may explain the MIL feature dominance. Importantly, the additional features allow for improved model interpretability. We demonstrate this in Figure 5.11 with two successful predictions, one in which the patient received steroids and one which did not. We observe substantial differences between the features of the two cases (e.g., age difference, difference in ratio of disease volume to total lung volume) and gain intuition for why the model successfully reached the prediction decision.

In all, we have demonstrated the potential for machine learning technology to evaluate CT scans acquired from patients diagnosed with COVID-19 and predict both pre-treatment and mid-treatment decisions related to steroid administration. In the CS1 analysis, the simple MIL technique achieved strong performance with an AUC of 0.85 +/- 0.10 while also providing temporal predictions that matched clinical expectations (e.g., the fit curves were separate upon hospitalization and converged nearing end of hospitalization). However, this performance was not maintained on a larger dataset, thus we investigated potential improvements to the model through improved feature extraction, pooling, and fusion techniques. Our findings failed to demonstrate significant improvement with these additions but

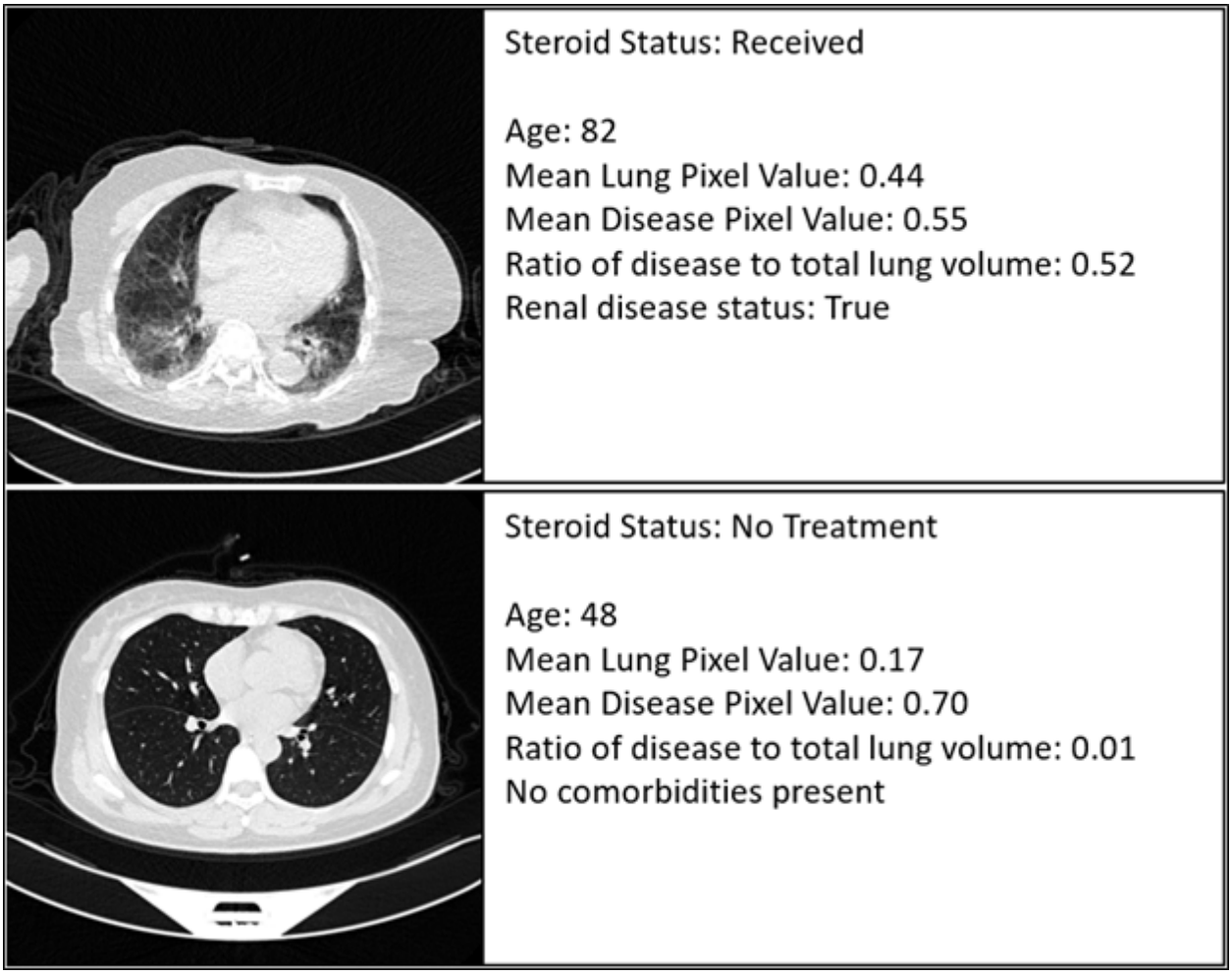

Figure 5.11: Example of two cases with successful prediction. (top) Patient received steroids. Investigating features reveals potentially contributing factors, including age, relatively small difference between mean disease and mean lung value, and large relative volume of disease. (bottom) Patient did not receive steroids, and based on the feature characteristics of extremely little diseased tissue volume and large difference between diseased and total lung pixel values, it becomes more clear why the model successfully reached the negative prediction decision.

provide a path for further investigations to improve our model. These methods may augment clinical decision-making and, ultimately, improve patient management and outcomes.

CHAPTER 6

SUMMARY AND FUTURE DIRECTIONS

In summary, the major contributions of this work are in the development of novel deep learning approaches to CT scan evaluation for the applications of CAC scoring and emphysema characterization on LDCT scans and for assessing COVID-19 patients leading to potential treatment recommendations and monitoring tools. These technologies could directly improve the quality of care and patient outcomes if successfully integrated as a component of the radiology clinical workflow for either lung screening protocols or severe COVID-19 patient evaluation, respectively.

When a high-risk patient is screened for lung cancer, a variety of other abnormalities can be visualized but would be time-consuming and tedious for a radiologist to review. Considering the typically heavy workload of radiologists, variability in subjective assessments, and potential errors that can be introduced by additional reading, automated tools to aid in the detection and diagnosis of non-lung cancer abnormalities are highly desirable.

In Chapter 3, we investigated automatic CAC scoring on LDCT scans, a task complicated by the lack of a standardized scoring system on screening scans and by other factors related to the LDCT acquisition, e.g., non-ECG gating, no contrast enhancement, and increased noise. The proposed CACU-Net can serve as a concurrent reader with a radiologist to reduce the variability in CAC score resulting from these complications. CACU-Net performs semantic segmentation on LDCT sections via a revised U-Net architecture with two branches, one focused on fine lesion information for accurate segmentation and one prioritizing coarser information regarding coronary artery branch location. With this model, we were able to demonstrate strong performance compared to other segmentation techniques both in the detection of CAC at the artery and case levels and in providing the ordinal CAC score. This was supported by ROC analysis, confusion matrices, and other error analyses.

In the future, there are several potential paths to investigate to improve the CACU-Net. In particular, the model optimization strategy did not target CAC scoring, the clinical end-goal, but was focused on accurate segmentation. While this initial strategy provided strong performance, it relies on human delineation of CAC lesions to serve as the reference standard; there is significant reader variability for reading CAC on LDCT scans, thus removing this bias as a major potential source of improvement and generalizability (although this would come at a reduction of score interpretability). Additionally, the error analysis found that many major model mistakes were caused by either lesion misclassification or by score inflation due to motion artifact. The impact of these could potentially be reduced by utilizing ensemble methods that were trained either on delineations provided by different radiologists or by combining multiple segmentation approaches that prioritize different information (e.g., one model focused on distinction between LMA and LAD lesions ensembled with another model focused on AVC rejection). And as with all deep learning studies, further testing on larger, more diverse, multi-institutional data would be a beneficial investigation to understand model generalizability. The inclusion of these investigations could potentially allow for more successful, generalizable CAC scoring on LDCT scans and may expand clinical options when significant CAC is discovered.

Chapter 4 investigated LDCT scans via deep attention-based multiple instance transfer learning to determine presence vs. absence of emphysema. This novel approach extracts quantitative feature information from individual LDCT sections then aggregates slice information to form a scan representation and classify between emphysema classes. Further, the attention weights used for slice aggregation were investigated as an interpretable model output through novel AWCs, which validated model performance based on expected trends between emphysema phenotypes and identified potentially important/confusing image features as identified by a radiologist. The combination of interpretable attention output, generalizability through transfer learning, and comparable performance to other published models,

the TAMIL approach demonstrated strong potential for clinical implementation, but further work must be completed prior to translation. As before, one of the most critical points of investigation is evaluation on a multi-institutional, diverse dataset. In addition, there have been more recent publications that utilize MIL such as the transformer-based approach investigated in Chapter 5; these may improve the performance for emphysema evaluation at the cost of increased parameterization. Finally, the model variants in this dissertation research were only trained for binary classification tasks, but more complicated training schemes including both unsupervised MIL and multi-class MIL for emphysema extent and phenotype may be beneficial investigations.

Finally, in Chapter 5, we investigated conventional and novel MIL schemes in the task of COVID-19 patient CT scan evaluation to provide pre-treatment and mid-treatment recommendations. A preliminary study found that a conventional MIL technique was able to achieve strong performance for both tasks on a small, limited dataset but failed to generalize to out-of-distribution data in the larger, more diverse database. Thus, we investigated three potential improvements to the cascaded transfer learning algorithm including the use of RadImageNet for transfer learning CT slice feature extraction, the use of transformers as an MIL pooling mechanism, and the use of multi-modal features. While the RadImageNet and transformer techniques failed to improve performance for the COVID-19 steroids administration classification task, we were able to evaluate the impact of different types of features and identify potential improvements for the future.

In the future, to augment this study, it would be clinically relevant to go beyond the initial classification task of steroid treatment initiation by predicting which patients are likely to respond/benefit from steroid treatments. The current model was trained with the assumption that the clinician's decision to treat was correct, a potentially inaccurate assumption with limited applicability; incorporation of patient responsiveness and outcome would significantly improve clinical utility but would require objective definition of patient

condition (e.g., Is a patient's response to treatment limited to visual regression of COVID-19 infection on imaging? Should additional clinical metrics be taken into account?). Similarly, the interpretable model output may be able to aid in identifying which imaging features are associated with good/poor outcomes to improve human-engineered feature production and, thus, multi-modal models. Finally, new advances in multi-modal approaches including representational gradient boosting may improve performance, particularly investigating the interaction between the different feature types.

In conclusion, this work demonstrates the strong potential for AI algorithms in CT scan evaluation including CAC scoring and emphysema characterization on lung cancer screening scans and COVID-19 treatment evaluations via diagnostic CT scans. For each use case, the AI approach achieved strong performance that suggests potential translation to the clinic following further validation. Overall, the development of novel deep learning architectures in this dissertation for a variety of use cases has demonstrated the potential for automated image reading for CT scan analysis; this technology may ultimately enhance clinical radiology workflow and improve patient care.

REFERENCES

[1] Jordan D. Fuhrman, Naveena Gorre, Qiyuan Hu, Hui Li, Issam El Naqa, and Maryellen L. Giger. A review of explainable and interpretable AI with applications in COVID-19 imaging. *Medical Physics*, 49(1):1–14, 2022. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/mp.15359.

[2] David Gunning. Explainable Artificial Intelligence (XAI). *machine learning*, page 18.

[3] Luke Danaher. Pulmonary emphysema | Radiology Reference Article | Radiopaedia.org.

[4] Natalia Antropova, Benjamin Q. Huynh, and Maryellen L. Giger. A deep feature fusion methodology for breast cancer diagnosis demonstrated on three imaging modality datasets. *Medical Physics*, 44(10):5162–5171, 2017. Number: 10 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/mp.12453.

[5] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based Deep Multiple Instance Learning. In *International Conference on Machine Learning*, pages 2127–2136. PMLR, July 2018. ISSN: 2640-3498.

[6] Thorsten M. Buzug. Computed Tomography. In Rüdiger Kramme, Klaus-Peter Hoffmann, and Robert S. Pozos, editors, *Springer Handbook of Medical Technology*, Springer Handbooks, pages 311–342. Springer, Berlin, Heidelberg, 2011.

[7] Jiang Hsieh. *Computed Tomography: Principles, Design, Artifacts, and Recent Advances*. SPIE Press, 2003. Google-Books-ID: JX__lLLXFHkC.

[8] Willi A Kalender. X-ray computed tomography. *Physics in Medicine and Biology*, 51(13):R29–R43, July 2006.

[9] Godfrey N. Hounsfield. Computed Medical Imaging. *Science*, 210(4465):22–28, 1980. Publisher: American Association for the Advancement of Science.

[10] A. M. Cormack. Representation of a Function by Its Line Integrals, with Some Radiological Applications. *Journal of Applied Physics*, 34(9):2722–2727, September 1963. Publisher: American Institute of Physics.

[11] A. M. Cormack. Representation of a Function by Its Line Integrals, with Some Radiological Applications. II. *Journal of Applied Physics*, 35(10):2908–2913, October 1964. Publisher: American Institute of Physics.

[12] Jiang Hsieh and Thomas Flohr. Computed tomography recent history and future perspectives. *Journal of Medical Imaging*, 8(5):052109, August 2021. Publisher: SPIE.

[13] Maryellen L. Giger. Machine Learning in Medical Imaging. *Journal of the American College of Radiology*, 15(3, Part B):512–520, March 2018. Number: 3, Part B.

[14] Berkman Sahiner, Aria Pezeshk, Lubomir M. Hadjiiski, Xiaosong Wang, Karen Drukker, Kenny H. Cha, Ronald M. Summers, and Maryellen L. Giger. Deep learning in medical imaging and radiation therapy. *Medical Physics*, 46(1):e1–e36, 2019. Number: 1 _eprint: https://aapm.onlinelibrary.wiley.com/doi/pdf/10.1002/mp.13264.

[15] Alvin Rajkomar, Jeffrey Dean, and Isaac Kohane. Machine Learning in Medicine. *New England Journal of Medicine*, 380(14):1347–1358, April 2019. Number: 14 Publisher: Massachusetts Medical Society _eprint: https://www.nejm.org/doi/pdf/10.1056/NEJMra1814259.

[16] Michael Chary, Saumil Parikh, Alex F. Manini, Edward W. Boyer, and Michael Radeos. A Review of Natural Language Processing in Medical Education. *Western Journal of Emergency Medicine*, 20(1):78–86, January 2019.

[17] Fei Wang, Lawrence Peter Casalino, and Dhruv Khullar. Deep Learning in Medicine—Promise, Progress, and Challenges. *JAMA Internal Medicine*, 179(3):293–294, March 2019.

[18] Effy Vayena, Alessandro Blasimme, and I. Glenn Cohen. Machine learning in medicine: Addressing ethical challenges. *PLOS Medicine*, 15(11):e1002689, November 2018. Publisher: Public Library of Science.

[19] James H. Thrall, Xiang Li, Quanzheng Li, Cinthia Cruz, Synho Do, Keith Dreyer, and James Brink. Artificial Intelligence and Machine Learning in Radiology: Opportunities, Challenges, Pitfalls, and Criteria for Success. *Journal of the American College of Radiology*, 15(3, Part B):504–508, March 2018.

[20] Garry Choy, Omid Khalilzadeh, Mark Michalski, Synho Do, Anthony E. Samir, Oleg S. Pianykh, J. Raymond Geis, Pari V. Pandharipande, James A. Brink, and Keith J. Dreyer. Current Applications and Future Impact of Machine Learning in Radiology. *Radiology*, 288(2):318–328, June 2018. Number: 2 Publisher: Radiological Society of North America.

[21] Charles J. Lynch and Conor Liston. New machine-learning technologies for computer-aided diagnosis. *Nature Medicine*, 24(9):1304–1305, September 2018. Number: 9 Publisher: Nature Publishing Group.

[22] Abraham Sathya, R A. Comparison of Supervised and Unsupervised Learning Algorithms for Pattern Classification. *International Journal of Advanced Research in Artificial Intelligence*, 2(2), 2013.

[23] Amir Atiya. Learning on a General Network. In D. Anderson, editor, *Neural Information Processing Systems*. American Institute of Physics, 1988.

[24] Yingjie Tian and Yuqi Zhang. A comprehensive survey on regularization strategies in machine learning. *Information Fusion*, 80:146–166, April 2022.

[25] Anouk Suppers, Alain J. van Gool, and Hans JCT Wessels. Integrated chemometrics and statistics to drive successful proteomics biomarker discovery. *Proteomes*, 6(2):20, 2018. Number: 2 Publisher: Multidisciplinary Digital Publishing Institute.

[26] Tulika Singh, Adarsh Ghosh, and Niranjan Khandelwal. Dimensional Reduction and Feature Selection: Principal Component Analysis for Data Mining. *Radiology*, 285(3):1055–1056, November 2017. Number: 3 Publisher: Radiological Society of North America.

[27] Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. Transfusion: Understanding Transfer Learning for Medical Imaging. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[28] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition, April 2015. arXiv:1409.1556 [cs].

[29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, Las Vegas, NV, USA, June 2016. IEEE.

[30] Amitojdeep Singh, Sourya Sengupta, and Vasudevan Lakshminarayanan. Explainable deep learning models in medical image analysis. *arXiv:2005.13799 [cs, eess]*, May 2020. arXiv: 2005.13799.

[31] Andreas Holzinger, Chris Biemann, Constantinos S. Pattichis, and Douglas B. Kell. What do we need to build explainable AI systems for the medical domain? *arXiv:1712.09923 [cs, stat]*, December 2017. arXiv: 1712.09923.

[32] Rohan Shad, John P. Cunningham, Euan A. Ashley, Curtis P. Langlotz, and William Hiesinger. Medical Imaging and Machine Learning. *arXiv:2103.01938 [cs, eess]*, March 2021. arXiv: 2103.01938.

[33] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, May 2019. Number: 5 Publisher: Nature Publishing Group.

[34] Cynthia Rudin and Joanna Radin. Why Are We Using Black Box Models in AI When We Don't Need To? A Lesson From An Explainable AI Competition. *Harvard Data Science Review*, 1(2), November 2019. Number: 2 Publisher: PubPub.

[35] Meherwar Fatima and Maruf Pasha. Survey of Machine Learning Algorithms for Disease Diagnostic. *Journal of Intelligent Learning Systems and Applications*, 09(01):1, 2017. Number: 01 Publisher: Scientific Research Publishing.

[36] Alvin Rajkomar, Jeffrey Dean, and Isaac Kohane. Machine Learning in Medicine. *New England Journal of Medicine*, 380(14):1347–1358, April 2019. Publisher: Massachusetts Medical Society _eprint: https://www.nejm.org/doi/pdf/10.1056/NEJMra1814259.

[37] Juri Yanase and Evangelos Triantaphyllou. A systematic survey of computer-aided diagnosis in medicine: Past and present developments. *Expert Systems with Applications*, 138:112821, December 2019.

[38] Di lin, Athanasios V. Vasilakos, Yu Tang, and Yuanzhe Yao. Neural networks for computer-aided diagnosis in medicine: A review. *Neurocomputing*, 216:700–708, December 2016.

[39] Nisreen I. R. Yassin, Shaimaa Omran, Enas M. F. El Houby, and Hemat Allam. Machine learning techniques for breast cancer computer aided diagnosis using different image modalities: A systematic review. *Computer Methods and Programs in Biomedicine*, 156:25–45, March 2018.

[40] Howard Lee and Yi-Ping Phoebe Chen. Image based computer aided diagnosis system for cancer detection. *Expert Systems with Applications*, 42(12):5356–5365, July 2015. Number: 12.

[41] Omer F Ahmad, Antonio S Soares, Evangelos Mazomenos, Patrick Brandao, Roser Vega, Edward Seward, Danail Stoyanov, Manish Chand, and Laurence B Lovat. Artificial intelligence and computer-aided diagnosis in colonoscopy: current evidence and future directions. *The Lancet Gastroenterology & Hepatology*, 4(1):71–80, January 2019. Number: 1.

[42] Morgan P. McBee, Omer A. Awan, Andrew T. Colucci, Comeron W. Ghobadi, Nadja Kadom, Akash P. Kansagra, Srini Tridandapani, and William F. Auffermann. Deep Learning in Radiology. *Academic Radiology*, 25(11):1472–1480, November 2018. Number: 11.

[43] Maciej A. Mazurowski, Mateusz Buda, Ashirbani Saha, and Mustafa R. Bashir. Deep learning in radiology: An overview of the concepts and a survey of the state of the art with focus on MRI. *Journal of Magnetic Resonance Imaging*, 49(4):939–954, 2019. Number: 4 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jmri.26534.

[44] Ying Song, Shuangjia Zheng, Liang Li, Xiang Zhang, Xiaodong Zhang, Ziwang Huang, Jianwen Chen, Ruixuan Wang, Huiying Zhao, Yunfei Zha, Jun Shen, Yutian Chong, and Yuedong Yang. Deep learning Enables Accurate Diagnosis of Novel Coronavirus (COVID-19) with CT images. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, pages 1–1, 2021. Conference Name: IEEE/ACM Transactions on Computational Biology and Bioinformatics.

[45] Nima Tajbakhsh, Laura Jeyaseelan, Qian Li, Jeffrey N. Chiang, Zhihao Wu, and Xiaowei Ding. Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. *Medical Image Analysis*, 63:101693, July 2020.

[46] Shervin Minaee, Yuri Y. Boykov, Fatih Porikli, Antonio J Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. Image Segmentation Using Deep Learning: A Survey. *IEEE*

Transactions on Pattern Analysis and Machine Intelligence, pages 1–1, 2021. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

[47] Adam Yala, Tal Schuster, Randy Miles, Regina Barzilay, and Constance Lehman. A Deep Learning Model to Triage Screening Mammograms: A Simulation Study. *Radiology*, 293(1):38–46, August 2019. Number: 1 Publisher: Radiological Society of North America.

[48] Mauro Annarumma, Samuel J. Withey, Robert J. Bakewell, Emanuele Pesce, Vicky Goh, and Giovanni Montana. Automated Triaging of Adult Chest Radiographs with Deep Artificial Neural Networks. *Radiology*, 291(1):196–202, January 2019. Number: 1 Publisher: Radiological Society of North America.

[49] Nicholas Bien, Pranav Rajpurkar, Robyn L. Ball, Jeremy Irvin, Allison Park, Erik Jones, Michael Bereket, Bhavik N. Patel, Kristen W. Yeom, Katie Shpanskaya, Safwan Halabi, Evan Zucker, Gary Fanton, Derek F. Amanatullah, Christopher F. Beaulieu, Geoffrey M. Riley, Russell J. Stewart, Francis G. Blankenberg, David B. Larson, Ricky H. Jones, Curtis P. Langlotz, Andrew Y. Ng, and Matthew P. Lungren. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of MRNet. *PLOS Medicine*, 15(11):e1002699, November 2018. Number: 11 Publisher: Public Library of Science.

[50] Hugo Jair Escalante, Sergio Escalera, Isabelle Guyon, Xavier Baró, Yağmur Güçlütürk, Umut Güçlü, and Marcel van Gerven, editors. *Explainable and Interpretable Models in Computer Vision and Machine Learning*. The Springer Series on Challenges in Machine Learning. Springer International Publishing, Cham, 2018.

[51] E. Tjoa and C. Guan. A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–21, 2020. Conference Name: IEEE Transactions on Neural Networks and Learning Systems.

[52] David Gunning, Mark Stefik, Jaesik Choi, Timothy Miller, Simone Stumpf, and Guang-Zhong Yang. XAI—Explainable artificial intelligence. *Science Robotics*, 4(37), December 2019. Number: 37 Publisher: Science Robotics Section: Focus.

[53] Feiyu Xu, Hans Uszkoreit, Yangzhou Du, Wei Fan, Dongyan Zhao, and Jun Zhu. Explainable AI: A Brief Survey on History, Research Areas, Approaches and Challenges. In Jie Tang, Min-Yen Kan, Dongyan Zhao, Sujian Li, and Hongying Zan, editors, *Natural Language Processing and Chinese Computing*, pages 563–574, Cham, 2019. Springer International Publishing.

[54] Xiao-Hui Li, Caleb Chen Cao, Yuhan Shi, Wei Bai, Han Gao, Luyu Qiu, Cong Wang, Yuanyuan Gao, Shenjia Zhang, Xun Xue, and Lei Chen. A Survey of Data-driven and Knowledge-aware eXplainable AI. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–1, 2020. Conference Name: IEEE Transactions on Knowledge and Data Engineering.

[55] Arun Das and Paul Rad. Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey. *arXiv:2006.11371 [cs]*, June 2020. arXiv: 2006.11371.

[56] Amina Adadi and Mohammed Berrada. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, 6:52138–52160, 2018. Conference Name: IEEE Access.

[57] Filip Karlo Došilović, Mario Brčić, and Nikica Hlupić. Explainable artificial intelligence: A survey. In *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pages 0210–0215, May 2018.

[58] Leihong Wu, Ruili Huang, Igor V. Tetko, Zhonghua Xia, Joshua Xu, and Weida Tong. Trade-off Predictivity and Explainability for Machine-Learning Powered Predictive Toxicology: An in-Depth Investigation with Tox21 Data Sets. *Chemical Research in Toxicology*, 34(2):541–549, February 2021. Number: 2 Publisher: American Chemical Society.

[59] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[60] CDC. Cases, Data, and Surveillance, February 2020.

[61] Edward F Patz, Erin Greco, Constantine Gatsonis, Paul Pinsky, Barnett S Kramer, and Denise R Aberle. Lung cancer incidence and mortality in National Lung Screening Trial participants who underwent low-dose CT prevalence screening: a retrospective cohort analysis of a randomised, multicentre, diagnostic screening trial. *The Lancet Oncology*, 17(5):590–599, May 2016. Number: 5.

[62] The National Lung Screening Trial Research Team. Reduced Lung-Cancer Mortality with Low-Dose Computed Tomographic Screening. *New England Journal of Medicine*, 365(5):395–409, August 2011. Number: 5.

[63] Claudia I Henschke, Dorothy I McCauley, David F Yankelevitz, David P Naidich, Georgeann McGuinness, Olli S Miettinen, Daniel M Libby, Mark W Pasmantier, June Koizumi, Nasser K Altorki, and James P Smith. Early Lung Cancer Action Project: overall design and findings from baseline screening. *The Lancet*, 354(9173):99–105, July 1999. Number: 9173.

[64] Peter J. Mazzone, Gerard A. Silvestri, Lesley H. Souter, Tanner J. Caverly, Jeffrey P. Kanne, Hormuzd A. Katki, Renda Soylemez Wiener, and Frank C. Detterbeck. Screening for Lung Cancer. *Chest*, 160(5):e427–e494, November 2021. Number: 5.

[65] Jyoti Malhotra, Matteo Malvezzi, Eva Negri, Carlo La Vecchia, and Paolo Boffetta. Risk factors for lung cancer worldwide. *European Respiratory Journal*, 48(3):889–902, September 2016. Publisher: European Respiratory Society Section: Series.

[66] Joanne T Chang, Gabriella M Anic, Brian L Rostron, Manju Tanwar, and Cindy M Chang. Cigarette Smoking Reduction and Health Risks: A Systematic Review and Meta-analysis. *Nicotine & Tobacco Research*, 23(4):635–642, April 2021.

[67] Nichole T. Tanner, Neeti M. Kanodra, Mulugeta Gebregziabher, Elizabeth Payne, Chanita Hughes Halbert, Graham W. Warren, Leonard E. Egede, and Gerard A. Silvestri. The Association between Smoking Abstinence and Mortality in the National Lung Screening Trial. *American Journal of Respiratory and Critical Care Medicine*, 193(5):534–541, March 2016. Number: 5.

[68] Nanda Horeweg, Ernst Th Scholten, Pim A de Jong, Carlijn M van der Aalst, Carla Weenink, Jan-Willem J Lammers, Kristiaan Nackaerts, Rozemarijn Vliegenthart, Kevin ten Haaf, Uraujh A Yousaf-Khan, Marjolein A Heuvelmans, Erik Thunnissen, Matthijs Oudkerk, Willem Mali, and Harry J de Koning. Detection of lung cancer through low-dose CT screening (NELSON): a prespecified analysis of screening test performance and interval cancers. *The Lancet Oncology*, 15(12):1342–1350, November 2014. Number: 12.

[69] Nikolaus Becker, Erna Motsch, Anke Trotter, Claus P. Heussel, Hendrik Dienemann, Philipp A. Schnabel, Hans-Ulrich Kauczor, Sandra González Maldonado, Anthony B. Miller, Rudolf Kaaks, and Stefan Delorme. Lung cancer mortality reduction by LDCT screening—Results from the randomized German LUSI trial. *International Journal of Cancer*, 146(6):1503–1513, March 2020. Number: 6.

[70] U. Pastorino, M. Silva, S. Sestini, F. Sabia, M. Boeri, A. Cantarutti, N. Sverzellati, G. Sozzi, G. Corrao, and A. Marchianò. Prolonged lung cancer screening reduced 10-year mortality in the MILD trial: new confirmation of lung cancer screening efficacy. *Annals of Oncology*, 30(7):1162–1169, July 2019. Number: 7.

[71] US Preventive Services Task Force, Alex H. Krist, Karina W. Davidson, Carol M. Mangione, Michael J. Barry, Michael Cabana, Aaron B. Caughey, Esa M. Davis, Katrina E. Donahue, Chyke A. Doubeni, Martha Kubik, C. Seth Landefeld, Li Li, Gbenga Ogedegbe, Douglas K. Owens, Lori Pbert, Michael Silverstein, James Stevermer, Chien-Wen Tseng, and John B. Wong. Screening for Lung Cancer: US Preventive Services Task Force Recommendation Statement. *JAMA*, 325(10):962, March 2021. Number: 10.

[72] Rowena Yip, Artit Jirapatnakul, Minxia Hu, Xiangmeng Chen, Dan Han, Teng Ma, Yeqing Zhu, Mary M. Salvatore, Laurie R. Margolies, David F. Yankelevitz, and Claudia I. Henschke. Added benefits of early detection of other diseases on low-dose CT screening. *Translational Lung Cancer Research*, 10(2):1141–1153, February 2021. Number: 2.

[73] Jessica González, Claudia I. Henschke, David F. Yankelevitz, Luis M. Seijo, Anthony P. Reeves, Rowena Yip, Yiting Xie, Michael Chung, Pablo Sánchez-Salcedo, Ana B. Alcaide, Aranzazu Campo, Juan Bertó, María del Mar Ocón, Jesus Pueyo, Gorka Bas-

tarrika, Juan P. de Torres, and Javier J. Zulueta. Emphysema phenotypes and lung cancer risk. *PLOS ONE*, 14(7):e0219187, July 2019. Number: 7 Publisher: Public Library of Science.

[74] Joseph Shemesh, Claudia I. Henschke, Dorith Shaham, Rowena Yip, Ali O. Farooqi, Matthew D. Cham, Dorothy I. McCauley, Mildred Chen, James P. Smith, Daniel M. Libby, Mark W. Pasmantier, and David F. Yankelevitz. Ordinal Scoring of Coronary Artery Calcifications on Low-Dose CT Scans of the Chest is Predictive of Death from Cardiovascular Disease. *Radiology*, 257(2):541–548, November 2010. Number: 2 Publisher: Radiological Society of North America.

[75] Giulia Veronesi, David R. Baldwin, Claudia I. Henschke, Simone Ghislandi, Sergio Iavicoli, Matthijs Oudkerk, Harry J. De Koning, Joseph Shemesh, John K. Field, Javier J. Zulueta, Denis Horgan, Lucia Fiestas Navarrete, Maurizio Valentino Infante, Pierluigi Novellis, Rachael L. Murray, Nir Peled, Cristiano Rampinelli, Gaetano Rocco, Witold Rzyman, Giorgio Vittorio Scagliotti, Martin C. Tammemagi, Luca Bertolaccini, Natthaya Triphuridet, Rowena Yip, Alexia Rossi, Suresh Senan, Giuseppe Ferrante, Kate Brain, Carlijn van der Aalst, Lorenzo Bonomo, Dario Consonni, Jan P. Van Meerbeeck, Patrick Maisonneuve, Silvia Novello, Anand Devaraj, Zaigham Saghir, and Giuseppe Pelosi. Recommendations for Implementing Lung Cancer Screening with Low-Dose Computed Tomography in Europe. *Cancers*, 12(6):1672, June 2020. Number: 6 Publisher: Multidisciplinary Digital Publishing Institute.

[76] Cristiano Rampinelli, Daniela Origgi, and Massimo Bellomi. Low-dose CT: technique, reading methods and image interpretation. *Cancer Imaging*, 12(3):548–556, February 2013.

[77] Arthur S. Agatston, Warren R. Janowitz, Frank J. Hildner, Noel R. Zusmer, Manuel Viamonte, and Robert Detrano. Quantification of coronary artery calcium using ultrafast computed tomography. *Journal of the American College of Cardiology*, 15(4):827–832, March 1990. Number: 4.

[78] Philip Greenland, Michael J. Blaha, Matthew J. Budoff, Raimund Erbel, and Karol E. Watson. Coronary Calcium Score and Cardiovascular Risk. *Journal of the American College of Cardiology*, 72(4):434–447, July 2018. Publisher: American College of Cardiology Foundation.

[79] Tamar S. Polonsky, Robyn L. McClelland, Neal W. Jorgensen, Diane E. Bild, Gregory L. Burke, Alan D. Guerci, and Philip Greenland. Coronary Artery Calcium Score and Risk Classification for Coronary Heart Disease Prediction. *JAMA*, 303(16):1610–1616, April 2010. Number: 16.

[80] Parveen K. Garg, Neal W. Jorgensen, Robyn L. McClelland, J. Adam Leigh, Philip Greenland, Michael J. Blaha, Andrew J. Yoon, Nathan D. Wong, Joseph Yeboah, and Matthew J. Budoff. Use of coronary artery calcium testing to improve coronary heart disease risk assessment in a lung cancer screening population: The Multi-Ethnic

Study of Atherosclerosis (MESA). *Journal of Cardiovascular Computed Tomography*, 12(6):493–499, November 2018. Number: 6.

[81] Joshua D. Mitchell, Nicole Fergestrom, Brian F. Gage, Robert Paisley, Patrick Moon, Eric Novak, Michael Cheezum, Leslee J. Shaw, and Todd C. Villines. Impact of Statins on Cardiovascular Outcomes Following Coronary Artery Calcium Scoring. *Journal of the American College of Cardiology*, 72(25):3233–3242, December 2018. Publisher: American College of Cardiology Foundation.

[82] Ahmed Bashir, William E. Moody, Nicola C. Edwards, Charles J. Ferro, Jonathan N. Townend, and Richard P. Steeds. Coronary Artery Calcium Assessment in CKD: Utility in Cardiovascular Disease Risk Assessment and Treatment? *American Journal of Kidney Diseases*, 65(6):937–948, June 2015.

[83] Cheryl D. Fryar, Te-Ching Chen, and Xianfen Li. Prevalence of uncontrolled risk factors for cardiovascular disease: United States, 1999-2010. *NCHS data brief*, (103):1–8, August 2012. Number: 103.

[84] Nagina Malguria, Stefan Zimmerman, and Elliot K. Fishman. Coronary Artery Calcium Scoring: Current Status and Review of Literature. *Journal of Computer Assisted Tomography*, 42(6):887–897, December 2018. Number: 6.

[85] Emily B. Tsai, Caroline Chiles, Brett W. Carter, Myrna C. B. Godoy, Girish S. Shroff, Reginald F. Munden, Mylene T. Truong, and Carol C. Wu. Incidental Findings on Lung Cancer Screening: Significance and Management. *Seminars in Ultrasound, CT and MRI*, 39(3):273–281, June 2018. Number: 3.

[86] Yu Htwe, Matthew D. Cham, Claudia I. Henschke, Harvey Hecht, Joseph Shemesh, Mingzhu Liang, Wei Tang, Artit Jirapatnakul, Rowena Yip, and David F. Yankelevitz. Coronary artery calcification on low-dose computed tomography: comparison of Agatston and Ordinal Scores. *Clinical Imaging*, 39(5):799–802, September 2015. Number: 5.

[87] W M Thurlbeck and N L Müller. Emphysema: definition, imaging, and quantification. *American Journal of Roentgenology*, 163(5):1017–1025, November 1994. Publisher: American Roentgen Ray Society.

[88] JOHN P. Wyatt, VERNON W. Fischer, and HERBERT C. Sweet. Panlobular Emphysema: Anatomy and Pathodynamics. *Diseases of the Chest*, 41(3):239–259, March 1962. Number: 3.

[89] A. E. Anderson and Alvan G. Foraker. Centrilobular emphysema and panlobular emphysema: two different diseases. *Thorax*, 28(5):547–550, September 1973. Number: 5 Publisher: BMJ Publishing Group Ltd Section: Articles.

[90] W L Foster, P C Pratt, V L Roggli, J D Godwin, R A Halvorsen, and C E Putman. Centrilobular emphysema: CT-pathologic correlation. *Radiology*, 159(1):27–32, April 1986. Number: 1 Publisher: Radiological Society of North America.

[91] Tetsuro Araki, Mizuki Nishino, Oscar E. Zazueta, Wei Gao, Josée Dupuis, Yuka Okajima, Jeanne C. Latourelle, Ivan O. Rosas, Takamichi Murakami, George T. O'Connor, George R. Washko, Gary M. Hunninghake, and Hiroto Hatabu. Paraseptal emphysema: Prevalence and distribution on CT and association with interstitial lung abnormalities. *European Journal of Radiology*, 84(7):1413–1418, July 2015. Number: 7.

[92] Javier J. Zulueta, Juan P. Wisnivesky, Claudia I. Henschke, Rowena Yip, Ali O. Farooqi, Dorothy I. McCauley, Mildred Chen, Daniel M. Libby, James P. Smith, Mark W. Pasmantier, and David F. Yankelevitz. Emphysema Scores Predict Death From COPD and Lung Cancer. *Chest*, 141(5):1216–1223, May 2012. Number: 5.

[93] Jurica Šprem, Bob D. de Vos, Nikolas Lessmann, Pim A. de Jong, Max A. Viergever, and Ivana Išgum. Impact of automatically detected motion artifacts on coronary calcium scoring in chest computed tomography. *Journal of Medical Imaging*, 5(4):044007, December 2018. Number: 4 Publisher: International Society for Optics and Photonics.

[94] Nikolas Lessmann, Bram van Ginneken, Majd Zreik, Pim A. de Jong, Bob D. de Vos, Max A. Viergever, and Ivana Išgum. Automatic Calcium Scoring in Low-Dose Chest CT Using Deep Neural Networks With Dilated Convolutions. *IEEE Transactions on Medical Imaging*, 37(2):615–625, February 2018. Number: 2 Conference Name: IEEE Transactions on Medical Imaging.

[95] Bob D. de Vos, Jelmer M. Wolterink, Tim Leiner, Pim A. de Jong, Nikolas Lessmann, and Ivana Išgum. Direct Automatic Coronary Calcium Scoring in Cardiac and Chest CT. *IEEE Transactions on Medical Imaging*, 38(9):2127–2138, September 2019. Number: 9 Conference Name: IEEE Transactions on Medical Imaging.

[96] Bob D. de Vos, Nikolas Lessmann, Pim A. de Jong, and Ivana Išgum. Deep Learning–Quantified Calcium Scores for Automatic Cardiovascular Mortality Prediction at Lung Screening Low-Dose CT. *Radiology: Cardiothoracic Imaging*, 3(2):e190219, April 2021. Number: 2 Publisher: Radiological Society of North America.

[97] Roman Zeleznik, Borek Foldyna, Parastou Eslami, Jakob Weiss, Ivanov Alexander, Jana Taron, Chintan Parmar, Raza M. Alvi, Dahlia Banerji, Mio Uno, Yasuka Kikuchi, Julia Karady, Lili Zhang, Jan-Erik Scholtz, Thomas Mayrhofer, Asya Lyass, Taylor F. Mahoney, Joseph M. Massaro, Ramachandran S. Vasan, Pamela S. Douglas, Udo Hoffmann, Michael T. Lu, and Hugo J. W. L. Aerts. Deep convolutional neural networks to predict cardiovascular risk from computed tomography. *Nature Communications*, 12(1):715, January 2021. Number: 1 Publisher: Nature Publishing Group.

[98] Carlos Cano-Espinosa, Germán González, George R. Washko, Miguel Cazorla, and Raúl San José Estépar. Automated Agatston score computation in non-ECG gated CT scans using deep learning. In *Medical Imaging 2018: Image Processing*, volume 10574, page 105742K. International Society for Optics and Photonics, March 2018.

[99] W. Wang, H. Wang, Q. Chen, Z. Zhou, R. Wang, H. Wang, N. Zhang, Y. Chen, Z. Sun, and L. Xu. Coronary artery calcium score quantification using a deep-learning algorithm. *Clinical Radiology*, 75(3):237.e11–237.e16, March 2020. Number: 3.

[100] Lea Azour, Michael A. Kadoch, Thomas J. Ward, Corey D. Eber, and Adam H. Jacobi. Estimation of cardiovascular risk on routine chest CT: Ordinal coronary artery calcium scoring as an accurate predictor of Agatston score ranges. *Journal of Cardiovascular Computed Tomography*, 11(1):8–15, January 2017. Number: 1.

[101] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. *arXiv:1505.04597 [cs]*, May 2015. arXiv: 1505.04597.

[102] Charles E. Metz and Xiaochuan Pan. "Proper" Binormal ROC Curves: Theory and Maximum-Likelihood Estimation. *Journal of Mathematical Psychology*, 43(1):1–33, March 1999. Number: 1.

[103] Elizabeth R. DeLong, David M. DeLong, and Daniel L. Clarke-Pearson. Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics*, 44(3):837–845, 1988. Number: 3 Publisher: [Wiley, International Biometric Society].

[104] A. E. Maxwell. Comparing the Classification of Subjects by Two Independent Judges. *The British Journal of Psychiatry*, 116(535):651–655, June 1970. Number: 535 Publisher: Cambridge University Press.

[105] Blaha Michael J., Mortensen Martin Bødtker, Kianoush Sina, Tota-Maharaj Rajesh, and Cainzos-Achirica Miguel. Coronary Artery Calcium Scoring. *JACC: Cardiovascular Imaging*, 10(8):923–937, August 2017. Number: 8 Publisher: American College of Cardiology Foundation.

[106] Yeqing Zhu, Rowena Yip, Joseph Shemesh, Artit C. Jirapatnakul, David F. Yankelevitz, and Claudia I. Henschke. Combined aortic valve and coronary artery calcifications in lung cancer screening as predictors of death from cardiovascular disease. *European Radiology*, 30(12):6847–6857, December 2020. Number: 12.

[107] Yeqing Zhu, Yong Wang, William E. Gioia, Rowena Yip, Artit C. Jirapatnakul, Michael S. Chung, David F. Yankelevitz, and Claudia I. Henschke. Visual scoring of aortic valve calcifications on low-dose CT in lung cancer screening. *European Radiology*, 30(5):2658–2668, May 2020. Number: 5.

[108] Sanne G. M. Van van Velzen, Bob D. de Vos, Julia M. H. Noothout, Helena M. Verkooijen, Max A. Viergever, and Ivana Išgum. Generative models for reproducible coronary calcium scoring. *Journal of Medical Imaging*, 9(5):052406, May 2022. Number: 5 Publisher: SPIE.

[109] Marc-André Carbonneau, Veronika Cheplygina, Eric Granger, and Ghyslain Gagnon. Multiple instance learning: A survey of problem characteristics and applications. *Pattern Recognition*, 77:329–353, May 2018.

[110] Xinggang Wang, Yongluan Yan, Peng Tang, Xiang Bai, and Wenyu Liu. Revisiting multiple instance neural networks. *Pattern Recognition*, 74:15–24, February 2018.

[111] Stephen M. Humphries, Aleena M. Notary, Juan Pablo Centeno, Matthew J. Strand, James D. Crapo, Edwin K. Silverman, and David A. Lynch. Deep Learning Enables Automatic Classification of Emphysema Pattern at CT. *Radiology*, 294(2):434–444, February 2020. Number: 2 Publisher: Radiological Society of North America.

[112] Andrea S. Oh, David Baraghoshi, David A. Lynch, Samuel Y. Ash, James D. Crapo, and Stephen M. Humphries. Emphysema Progression at CT by Deep Learning Predicts Functional Impairment and Mortality: Results from the COPDGene Study. *Radiology*, page 213054, May 2022. Publisher: Radiological Society of North America.

[113] Mohammadreza Negahdar, Adam Coy, and David Beymer. An End-to-End Deep Learning Pipeline for Emphysema Quantification Using Multi-label Learning. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 929–932, July 2019. ISSN: 1558-4615.

[114] Veronika Cheplygina, Lauge Sørensen, David M.J. Tax, Jesper Holst Pedersen, Marco Loog, and Marleen de Bruijne. Classification of COPD with Multiple Instance Learning. In *2014 22nd International Conference on Pattern Recognition*, pages 1508–1513, August 2014. ISSN: 1051-4651.

[115] S. N. Ørting, J. Petersen, L. H. Thomsen, M. M. W. Wille, and M. de Bruijne. Detecting emphysema with multiple instance learning. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 510–513, April 2018. ISSN: 1945-8452.

[116] Ruwan Tennakoon, Gerda Bortsova, Silas Ørting, Amirali K. Gostar, Mathilde M. W. Wille, Zaigham Saghir, Reza Hoseinnezhad, Marleen de Bruijne, and Alireza Bab-Hadiashar. Classification of Volumetric Images Using Multi-Instance Learning and Extreme Value Theorem. *IEEE Transactions on Medical Imaging*, 39(4):854–865, April 2020. Number: 4 Conference Name: IEEE Transactions on Medical Imaging.

[117] Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, and Yongbing Zhang. TransMIL: Transformer based Correlated Multiple Instance Learning for Whole Slide Image Classification. Technical Report arXiv:2106.00908, arXiv, October 2021. Issue: arXiv:2106.00908 arXiv:2106.00908 [cs] version: 2 type: article.

[118] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[119] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A. W. M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, December 2017.

[120] Zhaoyang Niu, Guoqiang Zhong, and Hui Yu. A review on the attention mechanism of deep learning. *Neurocomputing*, 452:48–62, September 2021.

[121] Coronavirus Disease (COVID-19) Situation Reports.

[122] Andrew Atkeson. What Will Be the Economic Impact of COVID-19 in the US? Rough Estimates of Disease Scenarios, March 2020.

[123] Ezekiel J. Emanuel, Govind Persad, Ross Upshur, Beatriz Thome, Michael Parker, Aaron Glickman, Cathy Zhang, Connor Boyle, Maxwell Smith, and James P. Phillips. Fair Allocation of Scarce Medical Resources in the Time of Covid-19. *New England Journal of Medicine*, 382(21):2049–2055, May 2020. Publisher: Massachusetts Medical Society _eprint: https://doi.org/10.1056/NEJMsb2005114.

[124] Zhe Xu, Lei Shi, Yijin Wang, Jiyuan Zhang, Lei Huang, Chao Zhang, Shuhong Liu, Peng Zhao, Hongxia Liu, Li Zhu, Yanhong Tai, Changqing Bai, Tingting Gao, Jinwen Song, Peng Xia, Jinghui Dong, Jingmin Zhao, and Fu-Sheng Wang. Pathological findings of COVID-19 associated with acute respiratory distress syndrome. *The Lancet Respiratory Medicine*, 8(4):420–422, April 2020. Publisher: Elsevier.

[125] Beth Russell, Charlotte Moss, Anne Rigg, and Mieke Van Hemelrijck. COVID-19 and treatment with NSAIDs and corticosteroids: should we be limiting their use in the clinical setting? *ecancermedicalscience*, 14:1023, March 2020.

[126] Dexamethasone in Hospitalized Patients with Covid-19. *New England Journal of Medicine*, 384(8):693–704, February 2021. Publisher: Massachusetts Medical Society _eprint: https://doi.org/10.1056/NEJMoa2021436.

[127] Yin Wang, Weiwei Jiang, Qi He, Cheng Wang, Baoju Wang, Pan Zhou, Nianguo Dong, and Qiaoxia Tong. A retrospective cohort study of methylprednisolone therapy in severe patients with COVID-19 pneumonia. *Signal Transduction and Targeted Therapy*, 5(1):1–3, April 2020. Number: 1 Publisher: Nature Publishing Group.

[128] Raef Fadel, Austin R Morrison, Amit Vahia, Zachary R Smith, Zohra Chaudhry, Pallavi Bhargava, Joseph Miller, Rachel M Kenney, George Alangaden, Mayur S Ramesh, and Henry Ford COVID-19 Management Task Force. Early Short-Course Corticosteroids in Hospitalized Patients With COVID-19. *Clinical Infectious Diseases*, 71(16):2114–2120, November 2020.

[129] Scott Simpson, Fernando U. Kay, Suhny Abbara, Sanjeev Bhalla, Jonathan H. Chung, Michael Chung, Travis S. Henry, Jeffrey P. Kanne, Seth Kligerman, Jane P. Ko, and Harold Litt. Radiological Society of North America Expert Consensus Document on

Reporting Chest CT Findings Related to COVID-19: Endorsed by the Society of Thoracic Radiology, the American College of Radiology, and RSNA. *Radiology: Cardiothoracic Imaging*, 2(2):e200152, March 2020. Number: 2 Publisher: Radiological Society of North America.

[130] Ming-Yen Ng, Elaine Y. P. Lee, Jin Yang, Fangfang Yang, Xia Li, Hongxia Wang, Macy Mei-sze Lui, Christine Shing-Yen Lo, Barry Leung, Pek-Lan Khong, Christopher Kim-Ming Hui, Kwok-yung Yuen, and Michael D. Kuo. Imaging Profile of the COVID-19 Infection: Radiologic Findings and Literature Review. *Radiology: Cardiothoracic Imaging*, 2(1):e200034, February 2020.

[131] Tao Ai, Zhenlu Yang, Hongyan Hou, Chenao Zhan, Chong Chen, Wenzhi Lv, Qian Tao, Ziyong Sun, and Liming Xia. Correlation of Chest CT and RT-PCR Testing in Coronavirus Disease 2019 (COVID-19) in China: A Report of 1014 Cases. *Radiology*, page 200642, February 2020.

[132] E. L. Kaplan and Paul Meier. Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, 53(282):457–481, June 1958. Publisher: Taylor & Francis _eprint: https://www.tandfonline.com/doi/pdf/10.1080/01621459.1958.10501452.

[133] W. Dana Flanders, Gary Tucker, Anusha Krishnadasan, Debra Martin, Eric Honig, and William M. McClellan. Validation of the pneumonia severity index. *Journal of General Internal Medicine*, 14(6):333–340, June 1999.

[134] N. Wu, J. Phang, J. Park, Y. Shen, Z. Huang, M. Zorin, S. Jastrzebski, T. Févry, J. Katsnelson, E. Kim, S. Wolfson, U. Parikh, S. Gaddam, L. L. Y. Lin, K. Ho, J. D. Weinstein, B. Reig, Y. Gao, H. Toth, K. Pysarenko, A. Lewin, J. Lee, K. Airola, E. Mema, S. Chung, E. Hwang, N. Samreen, S. G. Kim, L. Heacock, L. Moy, K. Cho, and K. J. Geras. Deep Neural Networks Improve Radiologists' Performance in Breast Cancer Screening. *IEEE Transactions on Medical Imaging*, 39(4):1184–1194, April 2020. Number: 4 Conference Name: IEEE Transactions on Medical Imaging.

[135] Fei Zhou, Ting Yu, Ronghui Du, Guohui Fan, Ying Liu, Zhibo Liu, Jie Xiang, Yeming Wang, Bin Song, Xiaoying Gu, Lulu Guan, Yuan Wei, Hui Li, Xudong Wu, Jiuyang Xu, Shengjin Tu, Yi Zhang, Hua Chen, and Bin Cao. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. *The Lancet*, 395(10229):1054–1062, March 2020.

[136] Michael J. Fine, Thomas E. Auble, Donald M. Yealy, Barbara H. Hanusa, Lisa A. Weissfeld, Daniel E. Singer, Christopher M. Coley, Thomas J. Marrie, and Wishwa N. Kapoor. A Prediction Rule to Identify Low-Risk Patients with Community-Acquired Pneumonia. *New England Journal of Medicine*, 336(4):243–250, January 1997. Publisher: Massachusetts Medical Society _eprint: https://doi.org/10.1056/NEJM199701233360402.

[137] Yi-Shan Lin, Wen-Chuan Lee, and Z. Berkay Celik. What Do You See? Evaluation of Explainable Artificial Intelligence (XAI) Interpretability through Neural Backdoors. *arXiv:2009.10639 [cs]*, September 2020. arXiv: 2009.10639.

[138] Nicholas P. Gruszauskas, Karen Drukker, Maryellen L. Giger, Ruey-Feng Chang, Charlene A. Sennett, Woo Kyung Moon, and Lorenzo L. Pesce. Breast US Computer-aided Diagnosis System: Robustness across Urban Populations in South Korea and the United States. *Radiology*, 253(3):661–671, December 2009. Publisher: Radiological Society of North America.

[139] Xueyan Mei, Zelong Liu, Philip M. Robson, Brett Marinelli, Mingqian Huang, Amish Doshi, Adam Jacobi, Chendi Cao, Katherine E. Link, Thomas Yang, Ying Wang, Hayit Greenspan, Timothy Deyer, Zahi A. Fayad, and Yang Yang. RadImageNet: An Open Radiologic Deep Learning Research Dataset for Effective Transfer Learning. *Radiology: Artificial Intelligence*, 4(5):e210315, September 2022. Publisher: Radiological Society of North America.

[140] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[141] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, and Sylvain Gelly. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[142] Xiaoyu Fang, Shen Li, Hao Yu, Penghao Wang, Yao Zhang, Zheng Chen, Yang Li, Liqing Cheng, Wenbin Li, Hong Jia, and Xiangyu Ma. Epidemiological, comorbidity factors with severity and prognosis of COVID-19: a systematic review and meta-analysis. *Aging (Albany NY)*, 12(13):12493–12503, July 2020.

[143] Adekunle Sanyaolu, Chuku Okorie, Aleksandra Marinkovic, Risha Patidar, Kokab Younis, Priyank Desai, Zaheeda Hosein, Inderbir Padda, Jasmine Mangat, and Mohsin Altaf. Comorbidity and its Impact on Patients with COVID-19. *SN Comprehensive Clinical Medicine*, 2(8):1069–1076, August 2020.

[144] Herve Abdi. Holm's Sequential Bonferroni Procedure. page 8.

[145] Yunyang Xiong, Zhanpeng Zeng, Rudrasis Chakraborty, Mingxing Tan, Glenn Fung, Yin Li, and Vikas Singh. Nyströmformer: A Nyström-based Algorithm for Approximating Self-Attention. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14138–14148, May 2021. Number: 16.

LIST OF PUBLICATIONS AND CONFERENCE PRESENTATIONS

PEER-REVIEWED PUBLICATIONS

JD Fuhrman, J Chen, Z Dong, FYM Lure, Z Luo, ML Giger. Cascaded deep transfer learning on thoracic CT in COVID-19 patients treated with steroids. *J. Med. Imaging.* **8**(S1), 014501 (2021), doi: 10.1117/1.JMI.8S1.014501

JD Fuhrman, N Gorre, Q Hu, H Li, I El Naqa, ML Giger. A review of explainable and interpretable AI with applications in COVID-19 imaging. *Med. Phys.* **49**:1-14 (2022), doi: 10.1002/mp.15359

A Mansour*, **JD Fuhrman***, F El Ammar, A Loggini, J Davis, C Lazaridis, C Kramer, FD Goldenberg, ML Giger. Machine learning for early detection of hypoxic ischemic brain injury after cardiac arrest. *Neurocrit. Care.* **6**:1-9 (2021), doi: 10.1007/s12028-021-0145-y

IM El Naqa, H Li, **JD Fuhrman**, Q Hu, N Gorre, W Chen, ML Giger. Lessons learned in transitioning to AI in the medical imaging of COVID-19. *J. Med. Imaging.* **8**(S1), 010902 (2021), doi: 10.1117/1.JMI.8.S1.010902

JD Fuhrman, R Yip, Y Zhu, AC Jirapatnakul, F Li, CI Henschke, DF Yankelevitz, ML Giger. Evaluation of emphysema on thoracic low-dose CTs through attention-based multiple instance deep learning. (under review)

JD Fuhrman, R Yip, Y Zhu, AC Jirapatnakul, F Li, DF Yankelevitz, CI Henschke, ML Giger. CACU-Net: Deep learning-based segmentation and location classification of coronary artery calcium in low-dose thoracic CT. (under review)

N Gorre, E Carranza, **JD Fuhrman**, ML Giger, H Li, IM El Naqa. CRP10 AI Interface - an integrated tool for exploring, testing and visualization of AI models. (under review)

JD Fuhrman*, C Wei*, EH Katsnelson*, BM Katsnelson*, H Li, F Li, Z Luo, Z Dong, FYM Lure, Z Cheng, ML Giger. Multi-modal deep learning of thoracic CT scans for COVID-19 patient prognosis and monitoring. (in draft)

EH Katsnelson*, C Wei*, BM Katsnelson*, **JD Fuhrman***, H Li, F Li, Z Luo, Z Dong, FYM Lure, ML Giger, Z Chung. COVID-19 Severity (SevScore) Based on Thoracic CT Scans by Patient Demographics and Comorbidities. (under review)

BM Katsnelson*, C Wei*, EH Katsnelson*, **JD Fuhrman***, H Li, F Li, Z Luo, Z Dong, FYM Lure, ML Giger, Z Cheng. (in draft)

CONFERENCE PRESENTATIONS AND POSTERS

JD Fuhrman, R Yip, Y Zhu, AC Jirapatnakul, DF Yankelevitz, CI Henschke, ML Giger. Multiple instance learning for automatic emphysema evaluation on low-dose CT lung cancer screening scans.

JD Fuhrman, R Yip, Y Zhu, AC Jirapatnakul, F Li, CI Henschke, DF Yankelevitz, ML Giger. Diagnosis of emphysema in low-dose CT screenings for lung cancer using multiple instance transfer learning. Oral Presentation at RSNA 2021.

JD Fuhrman, C Wei, F Li, H Li, Z Luo, Z Dong, FYM Lure, Z Cheng, ML Giger. Validation of deep transfer learning on CT scans for informing steroid treatment of 864 COVID-19 patients. Oral Presentation at RSNA 2021.

JD Fuhrman, L Kremer, R Yip, F Li, L Lan, H Li, AC Jirapatnakul, CI Henschke, DF Yankelevitz, ML Giger. Radiomic texture analysis for the assessment of osteoporosis on low-dose thoracic CT scans. Poster presentation at SPIE Medical Imaging 2021. Poster Presentation Award Honorable Mention.

JD Fuhrman, R Yip, AC Jirapatnakul, CI Henschke, DF Yankelevitz, ML Giger. A Machine Learning Pipeline for Segmentation and Classification of Coronary Artery Calcium Lesions in Lung Cancer Screening CT's. Featured Presentation at RSNA 2020. RSNA Trainee Research Prize Winner.

JD Fuhrman, Z Luo, J Chen, Y Su, M Ju, Z Dong, FYM Lure, ML Giger. Informing Steroid Administration for COVID-19 Patients Using Deep Learning. On-Demand Presentation at RSNA 2020.

JD Fuhrman, P Halloran, R Yip, AC Jirapatnakul, CI Henschke, DF Yankelevitz, ML Giger. Effect of observer variability and training cases on U-Net segmentation performance. Oral presentation at SPIE Medical Imaging 2020. Proceedings Paper: Volume 11316, Medical Imaging 2020: Image Perception, Observer Performance, and Technology Assessment, 113160T (2020) doi:10.1117/12.2549118

JD Fuhrman, R Yip, AC Jirapatnakul, CI Henschke, DF Yankelevitz, ML Giger. Cascade of U-Nets in the detection and classification of coronary artery calcium in thoracic low-dose CT. Oral presentation at SPIE Medical Imaging 2020. Proceedings Paper: Volume 11314, Medical Imaging 2020: Computer-Aided Diagnosis; 113140A (2020)

JD Fuhrman, R Yip, AC Jirapatnakul, CI Henschke, DF Yankelevitz, ML Giger. Deep learning in the task of detecting coronary artery calcifications on low-dose thoracic CTs. Snap oral presentation at the AAPM Annual Meeting 2019.

JD Fuhrman, R Yip, AC Jirapatnakul, CI Henschke, DF Yankelevitz, ML Giger. Detection and classification of coronary artery calcifications in low dose thoracic CT using deep learning. Poster presentation at SPIE Medical Imaging 2019. Proceedings Paper: Volume 10950, Medical Imaging 2019: Computer-Aided Diagnosis; 1095039 (2019)