CAPSULE NETWORK-BASED RADIOMICS: FROM DIAGNOSIS TO TREATMENT

Parnian Afshar

A thesis
in
The Department
of
Concordia Institute for Information Systems Engineering (CIISE)

Presented in Partial Fulfillment of the Requirements
For the Degree of Doctor of Philosophy
Concordia University
Montréal, Québec, Canada

August 2021

CONCORDIA UNIVERSITY
School of Graduate Studies

This is to certify that the thesis prepared

By: **Parnian Afshar**

Entitled: **Capsule Network-based Radiomics: From Diagnosis to Treatment**

and submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy (Information and Systems Engineering)

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the final examining commitee:

Dr. Sebastien Le Beux ______________________ Chair
Dr. Dimitri Androutsos ______________________ External Examiner
Dr. Amir Asif ______________________ External to Program
Dr. Abdessamad Ben Hamza ______________________ Examiner
Dr. Nizar Bouguila ______________________ Examiner
Dr. Arash Mohammadi ______________________ Supervisor

Approved ______________________________
Dr. Mohammad Mannan, Graduate Program Director

08/04/2021 ______________________________

Dr. Mourad Debbabi, Dean
Faculty of Engineering and Computer Science

Abstract

Capsule Network-based Radiomics: From Diagnosis to Treatment

Parnian Afshar, Ph.D.
Concordia University, 2021

Recent advancements in signal processing and machine learning coupled with developments of electronic medical record keeping in hospitals have resulted in a surge of significant interest in “radiomics". Radiomics is an emerging and relatively new research field, which refers to semi-quantitative and/or quantitative features extracted from medical images with the goal of developing predictive and/or prognostic models. Radiomics is expected to become a critical component for integration of image-derived information for personalized treatment in the near future. The conventional radiomics workflow is, typically, based on extracting pre-designed features (also referred to as hand-crafted or engineered features) from a segmented region of interest. Clinical application of hand-crafted radiomics is, however, limited by the fact that features are pre-defined and extracted without taking the desired outcome into account. The aforementioned drawback has motivated trends towards development of deep learning-based radiomics (also referred to as discovery radiomics). Discovery radiomics has the advantage of learning the desired features on its own in an and-to-end fashion. Discovery radiomics has several applications in disease prediction/ diagnosis. Through this Ph.D. thesis, we develop deep learning-based architectures to address the following critical challenges identified within the radiomics domain. First, we cover the tumor type classification problem, which is of high importance for treatment selection. We address this problem, by designing a Capsule network-based architecture that has several advantages over existing solutions such as eliminating the need for access to a huge amount of training data, and its capability to learn input transformations on its own. We apply different modifications to the Capsule network architecture to make it more suitable for radiomics. At one hand, we equip the proposed architecture with access to the tumor boundary box, and on the other hand, a multi-scale Capsule network architecture is designed. Furthermore, capitalizing on the advantages of ensemble learning paradigms, we design a boosting and also a mixture of experts capsule network. A Bayesian capsule network is also developed to capture the uncertainty of the tumor classification. Beside knowing the tumor type (through classification), predicting the patient's response to treatment plays an important role in treatment design. Predicting patient's response, including survival and tumor recurrence, is another goal of this thesis, which we address by designing a deep learning-based model that takes not only the medical images, but also different clinical factors (such as age and gender) as inputs. Finally, COVID-19 diagnosis, another challenging and crucial problem within the radiomics domain, is dealt with using both X-ray and Computed Tomography (CT) images (in particular low-dose ones), where two in-house datasets are collected for the latter and different capsule network-based models are developed for COVID-19 diagnosis.

Acknowledgments

First and foremost, I would like to thank my Supervisor, Dr. Arash Mohammadi, for his exceptional guidance and support, without which this thesis would not have been possible. I deeply express my gratitude to the committee members, Dr. Abdessamad Ben Hamza, Dr. Amir Asif, Dr. Nizar Bouguila, and Dr. Dimitri Androutsos, for evaluating this dissertation and their thoughtful feedback. I also want to thank our collaborators in University of Toronto, especially Prof. Plataniotis and Dr. Oikonomou, who have guided me through several steps of this thesis. I would like to thank my parents for their incredible encouragement, inspiration, and understanding, and last but not least, my husband whose unconditional support and patience helped me keep faith in myself.

Contents

List of Figures ix

List of Tables xiv

List of Abbreviations xv

List of CapsNet Symbols xvii

1 Overview of the Thesis 1

1.1 Introduction 1

1.2 Medical Resources for Radiomics 2

1.2.1 Screening Technologies 2

1.2.2 Complimentary Data Sources 4

1.3 Targeted Applications of Radiomics 4

1.4 Radiomics Challenges and Opportunities 5

1.5 Thesis Contributions 6

1.6 Organization of the Thesis 9

1.7 Publications 9

2 Literature Review 12

2.1 State-of-the-Art in Hand-Crafted Radiomics 12

2.1.1 Radiomics Feature Extraction 13

2.1.2 Radiomics Feature Reduction Techniques 17

2.1.3 Radiomics Statistical Analysis 20

2.1.4 Radiomics Stability 22

2.2 State-of-the-Art in Deep Learning-based Radiomics 24

2.2.1 Input Hierarchy 26

2.2.2 Pre-trained or Raw Models 27

2.2.3 Deep Learning Architectures in Radiomics 29

2.2.4 Explainability of Deep Learning-based Radiomics 41

2.3 Hybrid solutions to Radiomics 43

2.3.1 Combination of Radiomics and Other Data Sources 44

2.3.2 Fusion of HCR with DLR (i.e., Engineered Features Coupled with Deep Features) 46

2.4 Conclusion 49

3 Deep Learning-based Radiomics for Tumor Classification 50

3.1 Brain Tumor Classification 50

3.1.1 Brain Tumor Type Classification via Capsule Networks 51

3.1.2 BoostCaps: A Boosted Capsule Network for Brain Tumor Classification 57

3.1.3 BayesCap: A Bayesian Approach to Brain Tumor Classification Using Capsule Networks 61

3.2 Lung Nodule Classification 71

3.2.1 3D-MCN: 3D Multi-Scale Capsule Network for Lung Nodule Malignancy Classification 71

3.2.2 MIXCAPS: A Capsule Network-based Mixture of Experts for Lung Nodule Malignancy Prediction 80

3.3 Conclusion 93

4 Time-to-Event Outcome Prediction 95

4.1 Introduction 95

4.2 DRTOP Architecture based on Staging PET/CT Images 97

4.3 Experiments and Performance of the Proposed DRTOP 107

4.4 Conclusion 116

5 Deep Learning-based COVID-19 Diagnosis 118

5.1 Introduction 118

5.2 Capsule Networks for Identification of COVID-19 cases from Chest Radiographs (CXR) 120

5.3 Diagnosis of COVID-19 from CT scans 127

5.3.1 COVID-CT-MD, COVID-19 computed tomography scan dataset applicable in machine learning and deep learning 128

5.3.2 Hybrid Deep Learning Model for Diagnosis of COVID-19 using CT Scans and Clinical/Demographic Data 139

5.3.3 Human-level COVID-19 Diagnosis from Low-dose CT Scans Using a Two-stage Time-distributed Capsule Network 144

5.4 Conclusion 155

6 Conclusion and Future Direction 157

6.1 Summary of Contributions 157

6.2 Future Direction 160

List of Figures

1.1 Different Screening Technologies. 2

2.1 Extracting deep radiomics. The input to the network can be the original image, the segmented ROI, or the combination of both. Extracted radiomics features are either utilized through the rest of the network, or an external model is used to make the decision based on radiomics. 24

2.2 Taxonomy of deep learning-based radiomics (DLR). 26

2.3 Input hierarchy for one patient. In the top row the slice-level input is shown where the patients went through K examination visits during each of which $N^{(i)}$, for $(1 \leq i \leq K)$, number of slices is captured. The second row shows the volume-level where all slices associated with one visit is provided simultaneously as the input to the network. Finally, the third row shows the Patient-level analysis, where a single input consisting of all the volumes is provided. 27

2.4 Different angles of lung CT scan along with tumor crops in three different scales. 33

2.5 Capsule network architecture. 36

2.6 Routing by agreement. For the sake of simplicity number of output capsules is set to two. 38

2.7 Radiomics-based nomogram to predict lymph node metastasis. Tumor position is considered as extra information to assist with making a more reliable prediction. 46

2.8 Combining deep and hand-crafted Radiomics through feature-level or decision-level fusion. 47

3.1 Proposed Capsule network architecture for brain tumor classification. . . . 52

3.2 Capsnet and CNN accuracy for brain and segmented tumors images. 55

3.3 Defining the tumor boundary box. 55

3.4 The proposed BoostCaps framework. 58
3.5 ROC curve for the three tumor types predicted by the proposed BoostCaps. 60
3.6 ROC curve for the three tumor types predicted by the capsule network. . 61
3.7 Proposed BayesCap architecture, capable of outputting the mean prediction, as well as the prediction entropy. 64
3.8 Evolution of BayesCap and CNN weights mean and standard deviation. . . 68
3.9 Uncertainty histograms. 70
3.10 The proposed 3D-MCN framework. Three independent capsule networks take 3D nodule crops as inputs. Each CapsNet takes inputs at a different scale. The output vectors are masked and concatenated into a single vector. The resulted vector goes through a fusion module consisting of a set of fully connected layers to form probability associated with each class (benign or malignant). 72
3.11 Examples of available marked regions in the LIDC-IDRI dataset. Each regions can be classified as nodule or non-nodule. Nodules can also be categorized based on their size. Nodules larger than 3 *mm* are further grouped based on their malignancy ratings. 73
3.12 The ROC curve for the single-scale and multi-scale models, and the baseline model (i.e., fully connected layers trained on hand-crafted features only), showing that the proposed approach is capable of achieving higher AUC. . 75
3.13 Cases, where not all the single-scale models provided correct predictions. The check sign indicates the successful scale. This figure illustrates the necessity of including all the three scales in the final model. 77
3.14 Correlation between the CapsNet features and the hand-crafted features. Most of the features are positively or negatively correlated with volume and diameter. However, the correlation with x and y centers is not considerable, as the model is fed with cropped nodule slices in different scales, and the location with respect to the whole image is not accessible to the model. . . 77
3.15 T-SNE plot of the CapsNet learned features in 2D, showing that the features are capable of distinguishing between the two classes. 78
3.16 General framework of a mixture of experts approach. 81
3.17 Proposed MIXCAPS. 82

3.18 Example of nodules assigned to experts based on their volume and diameter. The nodule on the left, which has a lower probability of belonging to the first expert, is smaller in terms of volume and diameter compared to the nodule on the right. 89

3.19 Activation magnitude of the true class capsule, based on the tumor size and diameter. The darker the color, the higher the activation. (a) The true class capsule of the first expert has higher activation for larger nodules. (b) The true class capsule of the second expert has higher activation for smaller nodules. 89

3.20 MIXCAPS architecture with BoxCaps as experts for brain tumor type classification. 92

4.1 Proposed DRTOP model, where 3D CT and PET images go through separate networks, which are unsupervisedly pre-trained on an independent dataset. The outputs of the two networks (referred to as the CT risk and the PET risk) are combined with other clinical factors and fed to a Cox PHM. 97

4.2 The CT and PET components of the PET/CT, for patient 1 and patient 2, show the tumors in the superior segment of the left lower lobe and the right lower lobe respectively. CT and PET images for each patient are captured at the same level. 98

4.3 Frequency of number of tumor-containing slices, that can vary from 3 to 42. 100

4.4 Comparison between our proposed DRTOP model and hand-crafted radiomics, based on the concordance index. 109

4.5 Kaplan-Meier curves associated with the OS, with respect to (a) CT risk, (b) PET risk, and (C) age. Cut-off values to determine the low and high-risk groups are obtained from a logrank test. All predictors, when considered independently, have positive correlations with the OS. 110

4.6 Kaplan-Meier curves associated with the RFS, with respect to (a) PET risk, and (b) SUV. 110

4.7 Kaplan-Meier curves corresponding to (a) DC, and (b) LC, with respect to the CT risk, where cut-off values are determined by a logrank test. 111

4.8 One of the trees obtained from the random survival forest (RSF) to predict the recurrence free survival (RFS). Following this tree, one is able to obtain the cumulative hazard function (CHF), for each patient, at a desired time point. Based on the CHFs, the survival can also be calculated. 111

4.9 Correlation of the hand-crafted radiomics features extracted from PET images, with the deep learning-based features extracted from layer before the output layer of the model, trained on PET images. 113

4.10 Correlation of the hand-crafted radiomics features extracted from CT images, with the deep learning-based features extracted from layer before the output layer of the model, trained on CT images. 114

4.11 The CT and PET components cropped from three scales. These scales are stacked together and fed to the network. 115

5.1 The proposed COVID-CAPS architecture. 120

5.2 Labels available in the COVID-19 dataset. 124

5.3 ROC curve from the proposed COVID-CAPS. Without pre-training refers to training from scratch. 124

5.4 The distribution of the Exposure values for COVID-19, CAP and Normal cases. 131

5.5 (a) The number of cases separated by the patient's gender. (b) The distribution of age for COVID-19, CAP and Normal cases. 134

5.6 (a) The distribution of the Infection Ratio in the labeled dataset for COVID-19 and CAP cases. (b) The histogram of the Infection Ratio in the labeled dataset for COVID-19 and CAP cases. 135

5.7 Average Infection Ratio in each lobe of the lung for COVID-19 and CAP cases in the labeled dataset. 136

5.8 Structure of the data included in COVID-CT-MD dataset. 137

5.9 Pipeline of the proposed hybrid model for COVID-19 diagnosis. 139

5.10 The internal structure of one of the decision trees created by the RF component of the proposed hybrid model. CT refers to the probability score obtained from the CapsNet model. 143

5.11 The proposed 2 stage deep learning model for COVID-19 diagnosis using LDCT/ULDCT. At the first stage, CT slices go through a capsule network, one by one, to detect those with evidence of infection. At the second stage, 10 most probable slices with infection detected in the previous stage go through a time-distributed capsule network, output of which determines the probability of COVID-19, CAP, and normal, after applying a global max pooling. 147

5.12 The MLP model combining the output of the two-stage deep learning model with the clinical data. Clinical data includes demographic characteristics and 5 common COVID-19 and CAP symptoms. Four sets of fully connected layers determine the final output. 148

5.13 ROC curve for COVID-19 diagnosis (vs CAP and normal) using the proposed deep learning model and CT scans only. 149

5.14 Grad-CAM visualization of one CAP slice. This figure shows that the proposed AI model is paying attention to relevant locations of the image. . . . 151

5.15 Grad-CAM visualization of two COVID-19 slices. This figure shows that the proposed AI model is paying attention to relevant locations of the image. . 153

5.16 ROC curve for COVID-19 diagnosis (vs CAP and normal) using the proposed deep learning model and both CT scans and clinical data. 154

List of Tables

2.1 Different categories of HCR features commonly used within the context of radiomics. 14

2.2 Feature reduction techniques commonly used within the Radiomics literature. 18

2.3 Common analysis methods in Radiomics. 20

2.4 A Comparison between hand-Crafted and Deep Radiomics. 44

3.1 Brain tumor classification accuracy based on different capsule network architectures. 54

3.2 Training hyper-parameters used for brain tumor classification via Adam optimizer. 57

3.3 Comparison between the proposed CapsNet BoxCaps and previous results. The bold number corresponds to the proposed approach, which outperforms its counterparts. 57

3.4 Results obtained from the proposed BoostCaps and the capsule network. . 62

3.5 Performance of the proposed 3D-MCN approach along with performance of single scale-models, on the independent test set. The proposed approach outperforms others in terms of the AUC, accuracy and sensitivity. 75

3.6 Effect of changing the weights of the terms controlling the false positives and false negatives. As expected, it is observed that assigning more weight to the false positive loss, increases the sensitivity, while putting higher weight to the false negative loss, increases the specificity. 76

3.7 A list of papers that have used LIDC-IDRI to predict the lung nodule malignancy based on the radiologists' provided ratings. Our proposed approach outperforms all the others except the third one. However, the third model incorporates hand-crafted features, needing exact annotations from the radiologists. Our proposed model, however, only requires nodule rough boundary box. 79

3.8 Performance of the proposed MIXCAPS compared to that of a single capsule network, a mixture of CNNs and a single CNN. Numbers in parenthesis show the 95% confidence intervals. 86
3.9 Performance of the proposed MIXCAPS compared to that of an ensemble of capsule networks. Numbers in parenthesis show the 95% confidence intervals. 87
3.10 Performance of the proposed MIXCAPS with BoxCaps as experts. Numbers in parenthesis show the 95% confidence intervals. 91
4.1 Datasets used in the literature for lung cancer time-to-event outcome predictions. 101
4.2 Results from the proposed DRTOP model. HR stands for hazard ratio (exponent of the obtained coefficient). Significant predictors are obtained based on an F-test, with a significance level of 0.05. Concordance index is calculated on the test set. 104
4.3 Hand-crafted features, extracted in DRTOP study. 105
4.4 Results obtained from hand-crafted radiomics. As there is no significant predictor for the LC, Concordance index is not calculated. 108
4.5 Variable importance values obtained from the RSF, for recurrence free survival prediction. The negative value means no predictive importance. . . . 111
4.6 Results obtained from the proposed MDR-SURV and the single-scale methods. 112
5.1 Results obtained from the proposed COVID-CAPS, along with the results from Reference [170]. Pre-trained CNN refers to a CNN with the same front-end as the COVID-CAPS 123
5.2 Description of the External X-ray images dataset used for pre-training COVID-CAPS. 124
5.3 Available COVID-19 CT scan datasets. NA stands for not available. 128
5.4 CT scan settings used to acquire the COVID-CT-MD dataset. 131
5.5 The statistical parameters (mean and standard deviation) of the Exposure values. 131
5.6 Gender and age distribution in COVID-CT-MD 133
5.7 The number of cases, Slices, and Infection Ratio in the labeled dataset. . . 134
5.8 Number of cases and slices, respectively, demonstrating infection in each lobe. LLL: Left Lower Lobe – LUL: Left Upper Lobe – RLL: Right Lower Lobe and Lingula – RML: Right Middle Lobe – RUL: Right Upper Lobe . 135

5.9 Patient-level classification results obtained from the proposed Hybrid Model and its counterparts. 142

5.10 Features contributing the most to the final RF decision. 143

5.11 Characteristics of the in-house dataset. SD stands for standard deviation. . 148

5.12 Performance of the AI model and the radiologist blind to the labels, using only CT scans. 150

5.13 Performance of the AI model and the radiologist blind to the labels, using both CT scans and clinical data. 152

6.1 Summary of the contributions. 159

List of Abbreviations

AUC	Area under the curve
CAE	Convolutional auto-encoder
CAP	Community acquired pneumonia
CapsNet	Capsule Network
CHF	Cumulative hazard function
CNN	Convolutional neural network
CT	Computed tomography
DAE	Denoising auto-encoder
DBM	Deep Boltzmann machine
DBN	Deep belief network
DC	Distant control
DLR	Deep learning-based radiomics
DT	Decision tree
HCR	Hand-crafted radiomics
KMS	Kaplan-Meier survival
LC	Local control
LDCT	Low-dose CT
MRI	Magnetic resonance imaging
NN	Neural network
OS	Overall survival

PET	Positron emission tomography
PHM	Proportional hazards model
ROI	Region of inetrest
RT-PCR	Reverse transcription polymerase chain reaction
RBM	Restricted Boltzmann machine
RF	Random forest
RFS	Recurrence free survival
RSF	Random survival forest
RNN	Recurrent neural network
SUV	Standardized uptake value
SVM	Support vector machine
T-SNE	T-distributed stochastic neighbor embedding
ULDCT	Ultra-low-dose CT

List of CapsNet Symbols

$\boldsymbol{u}$	Capsule instantiation parameters vector
$\boldsymbol{W}$	Trainable weight matrix
$\hat{\boldsymbol{u}}$	Capsule prediction
a	Agreement
c	Prediction score, Coupling coefficient
$\boldsymbol{s}$	Capsule output
l	Margin loss
m^+	Hyper-parameter
m^-	Hyper-parameter
λ	Hyper-parameter

Chapter 1

Overview of the Thesis

1.1 Introduction

Radiomics [1–15] refers to the process of extracting and analyzing several semi-quantitative (e.g., attenuation, shape, size, and location) and/or quantitative features (e.g., histogram and grey-level intensity) from medical images with the ultimate goal of obtaining predictive or prognostic models. Although several challenges are in the way of bringing radiomics into daily clinical practice, it is expected that radiomics become a critical component for integration of image-driven information for personalized treatment in the near future.

The first comprehensive clinical application of radiomics [16–19] was performed by Aerts *et al.* [20] with involvement of $1,019$ lung cancer patients. More than 400 different intensity, shape, and texture features were extracted from Computed Tomography (CT) images and used together with clinical information and gene expression data to develop radiomics heat map, which shows the association between radiomics and different clinical outcomes such as cancer stage. This clinical study has illustrated/validated effectiveness of radiomics for tumor related predictions and showed that radiomics has the capability to identify lung and head-and-neck cancers from a single-time point CT scan. Consequently, there has been a surge of interest [21–25] on this multidisciplinary research area as radiomics has the potential to provide significant assistance for assessing the risk of recurrence of cancer [26]; Evaluating the risk of radiation-induced side-effects on non-cancer tissues [27], and; Predicting the risk for cancer development in healthy subjects [27].

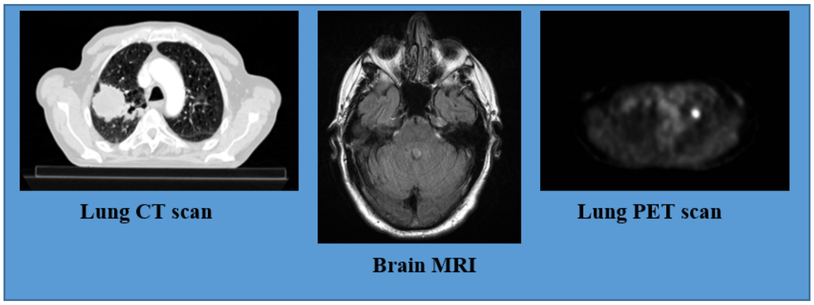


Figure 1.1: Different Screening Technologies.

The key underlying hypothesis in the radiomics is that the constructed descriptive models (based on medical imaging data, sometimes complemented by biological and/or medical data) are capable of providing relevant and beneficial predictive, prognostic, and/or diagnostic information. In this regard, one can identify two main categories of radiomics. Conventional pipeline based on Hand-Crafted radiomics features (HCR) that consists of the following four main processing tasks: (i) Image acquisition/reconstruction; (ii) Image segmentation; (iii) Feature extraction and quantification, and; (iv) Statistical analysis and model building. On the other hand, the Deep Learning-based radiomics (DLR) pipeline has recently emerged, which differs from the former category since deep networks do not necessarily need the segmented Region Of Interest (ROI), and their feature extraction and analysis parts are partially or fully coupled. Before highlighting radiomics challenges and contributions of the thesis, next, medical resources available for radiomics are briefly presented.

1.2 Medical Resources for Radiomics

Several potential medical resources provide information to the radiomics pipeline, some of which are directly used to extract radiomics features, while some serve the decision making process, as complementary information sets. Below, we briefly review the most important data resources for radiomics.

1.2.1 Screening Technologies

The radiomics features can be extracted from several imaging modalities, as shown in Fig. 1.1, among which the following are the most commonly used modalities:

- ***Computed Tomography (CT) Scans***: The CT is the modality of choice for the diagnosis of many diseases in different parts of the body, and by providing high resolution images [16] paves the path for extracting comparable radiomics features. Nonetheless, the CT imaging performance depends on different components of the utilized protocol including the following three main properties: (i) Slice thickness, which is the distance in millimetre (mm) between two consecutive slices; (ii) The capability for projecting the density variations into image intensities, and; (iii) Reconstruction algorithm, which aims at converting tomographic measurements to cross-sectional images. Although CT protocols for specific clinical indications are usually similar across different institutions, radiomics features can even differ between different scanners with the same settings [28]. Therefore, there is still a considerable need to ensure consistency of radiomics feature extraction amongst different scanners and imaging protocols [17]. CT images are typically divided into two categories [29]: screening and diagnostic. While screening CT uses low dose images, diagnostic CT utilizes high dose and is of higher quality and contrast.

- ***Positron Emission Tomography (PET) Scans***: The PET is a nuclear imaging modality that evaluates body function and metabolism [16], and since its performance depends on not only the scanner properties, but also the doze calibration, similar to the case with the CT scans, standardizing the PET protocols across different institutions is challenging. Furthermore, glucose level at the time of scanning can also affect the properties of PET images [17].

- ***Magnetic Resonance Imaging (MRI)***: Unlike CT, properties of MRI images are not directly associated with tissue density and specific methods are required to obtain the so-called signal intensity. Besides, several imager and vendor-dependant factors such as gradient and coil systems [30], pulse sequence design, slice thickness, and other parameters such as artifacts and magnetic field strength affect the properties of the MRI images [17], which should be consistent across different institutions.

1.2.2 Complimentary Data Sources

In addition to imaging resources, the following clinical data sources are typically combined with radiomics features:

- ***Gene Expression***: The process of converting DNA to functional product to have a global insight of cellular function.
- ***Clinical Characteristics***: Patient's characteristics such age, gender, and past medical and family history [17].
- ***Blood Bio-markers***: Measurable characteristics from the patient's blood such as glucose level, cholesterol level and blood pressure.
- ***Prognostic Markers***: Markers to evaluate the progress of the disease, response to treatment or survival, such as size, tumor stage, tumor recurrence, and metastasis.

1.3 Targeted Applications of Radiomics

In recent years, radiomics has been applied to several health-care applications, including oncology, cardiology, and neurology. In cardiology, for instance, radiomics is used in different investigations, such as identifying the coronary plaques [31]. In neurology, it is widely applicable for detecting Alzheimer's disease [32] and Parkinson's disease [33]. However, among all the applications of the radiomics, cancer-related topics, such as diagnosis, detection, classification, and survival and recurrence prediction, have been the focus of interest. Furthermore, in the recent year, the COVID-19 pandemic and its consequences have caused a trend towards exploring radiomics for the diagnosis of COVID-19 using chest radiographs, having the promise of compensating for the low sensitivity and accessibility of the Reverse Transcription Polymerase Chain Reaction (RT-PCR) test.

In particular, the thesis covers three main applications of the radiomics:

- ***Tumor Classification***, which refers to determining the type of the tumor. Typically, cancer is classified into the following main classes: (i) benign; (ii) primary malignant, and; (iii) metastatic malignant, based on several factors such as their ability to spread to other tissues. Benign tumors usually do not spread to other organs but may need surgical resection because occasionally they may grow in size. Pre-invasive lesions may be indolent for years, however, they may transform to aggressive malignant tumors and therefore need to be monitored closely or even be treated with lower dose of anti-cancer regimens. Malignant tumors are life threatening and may spread to distant organs, requiring more complicated treatments such as Chemotherapy. Prediction of tumor malignancy likelihood with noninvasive methods such as Radiomics is, therefore, of paramount importance.

- ***Time-to-event outcome Prediction***: The knowledge of the expected survival of a specific disease with or without a specific treatment is critical both for physicians and the patients. Physicians need to choose the best treatment plan for their patients and patients need to know their predicted survival time in order to make their own choices for the quality of their life. Radiomics can add significant information about patient's survival based on image properties and heterogeneity of the tumor and this has attracted a lot of attention recently.

- ***COVID-19 Diagnosis***: Since the beginning of the coronavirus disease (COVID-19) outbreak in December 2019 in Wuhan, China, a global healthcare crisis has emerged. Currently, RT-PCR is considered as the gold standard method in COVID-19 diagnosis. RT-PCR is, however, prone to a number of limitations, i.e., besides being time consuming, it is associated with high false-negative rate in different clinical samples. Due to high sensitivity and rapid access, chest CT scan has been the main imaging modality for diagnosis, prognostic assessment, and detection of complications of COVID-19. CT scan, by means of radiomics, can contribute to assessing the complications, extent of COVID-19 involvement, and risk of intensive care unit (ICU) admission.

1.4 Radiomics Challenges and Opportunities

Despite recent advancements in the field of radiomics and increase of its potential clinical applications, there are still several open problems, which require extensive investigations, including:

C1. Most radiomics models need rich amounts of training images, however, due to strict privacy policies, medical images are usually hard to collect.

C2. Even without considering the privacy issues, it is difficult to find the required amount of data with similar clinical characteristics (e.g., corresponding to the same cancer stage).

C3. Radiomics analysis need ground truth, which is scarce as labels can only be provided by clinical experts (this is in contrary to other multimedia domains). This calls for development of weakly or semi-supervised solutions taking into account the specifics of the radiomics domain.

C4. Unbalanced data refers to the problem where classes are not equal in a classification task rendering the classifier biased toward the majority class. This is almost all of the time the case for radiomics analysis as the number of positive classes (existence of disease) is typically smaller than the negative ones. Therefore, proper care is needed when working with medical data. Although several solutions, such as modifying the metric function to give more weight to minority class, are provided to deal with the aforementioned issue, it is still an unsolved problem that needs further investigations.

C5. The biggest challenge in combining various data sources (such as imaging and clinical) is that not all data is provided for all the patients. In other words, radiomics analysis model should be equipped with the ability to work with sparse data. Besides, the currently used fusion strategies within the radiomics are still in their infancy and development of more rigorous fusion rules is necessary. For instance, feature-level fusion results in a vector and how to sort/combine the localized feature vectors is an open challenge. Giving the superiority of the initial results obtained from hybrid radiomics, this issue becomes an urgent matter calling for advanced mixture of expert solutions.

1.5 Thesis Contributions

Below, the contributions of the thesis are briefly outlined:

- **Chapter 3: Deep Learning-based Radiomics for Tumor Classification**

1. A capsule network-based brain tumor classification is developed [2, 3], which benefits from the capability of this kind of architecture in handling small datasets, targeting challenges C1 and C2, as introduced in Section 1.4. Capsule networks, however, tend to account for every detail present in the input, and as such, they perform better when fed with segmented tumors rather than the whole brain tissue. Nevertheless, segmenting the tumors is time-consuming and burdensome, which led us to develop a modified version of the capsule network, which is fed with rough tumor boundaries. In other words, this framework does not require fine annotations, and the model has access to both brain tissues and location of the tumor.

2. A boosted capsule network, referred to as **BoostCaps** [4] is proposed that benefits from boosting to gradually enhance an initial weak learner (a simple capsule network). This framework does not require a full exploration within the domain of all possible CapsNet architectures. In fact, the algorithm starts with a simple design and by giving more weight to misclassified samples, the model improves itself through steps.

3. **BayesCap** [5], a Bayesian capsule network, which is designed to account for uncertainty in the model's weights. The output of BayesCap is not only the average over predictions (through Monte-Carlo simulation), but also entropy as a measure of uncertainty with the promise of returning uncertain prediction to human experts.

4. To increase the accuracy of lung nodule malignancy prediction, a 3D multiscale capsule network, referred to as **3D-MCN** [6], is proposed. This model is fed with multi-scale tumor crops, with the intuition that the nodule morphological characteristics are not the only indicators of its malignancy, and incorporation of information obtained from the surrounding tissues and vessels play a critical role in determining the type of the nodule.

5. Capitalizing on the success of ensemble models, a mixture of capsule networks (**MIXCAPS** [7]) is designed. Each expert is shown to be focusing on particular nodules, and the final output is the weighted average over all the predictions. It is also shown that capsule networks, themselves, can be viewed as mixture of experts, and thus, the proposed model is a

hierarchical mixture of experts. This contribution targets challenge C5 in Section 1.4.

- **Chapter 4: Time-to-Event Outcome Prediction [8, 9]**

  1. A novel deep learning architecture, referred to as **DRTOP** is possessed to predict pre-defined clinical endpoints in a cohort of lung cancer patients before the initiation of treatment, based on staging PET/CT images. The proposed DRTOP model considers not only PET and CT images, but also clinical factors, such as age and gender. The model is pre-trained as a convolutional auto-encoder on an external unlabeled dataset, in an unsupervised manner, targeting challenge C3 in Section 1.4. Cox proportional hazards model (PHM) and random survival forest (RSF) are trained on the output of the deep learning model. The features extracted from the deep model show high interpretability when compared with hand-crafted features.

- **Chapter 5: Deep Learning-based COVID-19 Diagnosis**

  1. To tackle the crucial problem of COVID-19 diagnosis, a capsule network (referred to as **COVID-CAPS** [10]) is designed, which is fed with X-ray images of COVID-19, normal, and pneumonia cases. The model is pre-trained on an external X-ray dataset, and loss function is modified to account for class imbalance, as discussed in challenge C4 in Section 1.4.
  2. CT scans are more sensitive compared to X-ray images as they capture the tissue 3D characteristics. In this sense, an in-house dataset of standard-dose CT scans (referred to as **COVID-CT-MD** [11]), along with clinical information is collected and used in designing a hybrid deep learning model [12].
  3. Standard-dose CT scans have the disadvantage of high radiation exposure. Therefore, first a dataset of low-dose and ultra-low-dose CT scans (LDCT and ULDCT) is collected, followed by designing a time-distributed deep learning model, analyzing the slices simultaneously [13], where the final decision is a weighted average over the prediction coming from single slices. Experiments show the proposed model achieves human-level performance.

1.6 Organization of the Thesis

Chapter 1 (this chapter) provided an overview and a summary of important contributions made in the thesis. As stated previously, through this Ph.D. thesis, we cover three main applications of the radiomics, i.e., tumor classification, time-to-event outcome prediction, and COVID-19 diagnosis, with focus on developing deep learning-based architectures. These topics will be covered through the following chapters:

- **Chapter 2** provides the literature review of the topic.
- In **Chapter 3**, we concentrate on the deep learning-based solutions to tumor classification.
- Time-to-event outcome (in particular survival) prediction is investigated in **Chapter 4**.
- Radiomics-based COVID-19 diagnosis is provided in **Chapter 5**.
- In **Chapter 6**, we conclude the thesis and future direction will be discussed.

1.7 Publications

Journal Publications

[J8] **P. Afshar**, M. J. Rafiee, F. Naderkhani, Sh. Heidarian, N. Enshaei, A. Oikonomou, F. Babaki Fard, R. Anconina, K. N. Plataniotis, K. Farahani, A. Mohammadi,“Human-level COVID-19 Diagnosis from Low-dose CT Scans Using a Two-stage Time-distributed Capsule Network,” *arXiv:2105.14656v1*, 2021.

[J7] **P. Afshar**, Sh. Heidarian, N. Enshaei, F. Naderkhani, M. J. Rafiee, A. Oikonomou, F. Babaki Fard, K. Samimi, K. N. Plataniotis, A. Mohammadi, “COVID-CT-MD: COVID-19 Computed Tomography (CT) Scan Dataset Applicable in Machine Learning and Deep Learning,” *Nature Scientific Data*, vol 8, 121, 2021.

[J6] **P. Afshar**, F. Naderkhani, A. Oikonomou, M. J. Rafiee, A. Mohammadi, K. N. Plataniotis, “MIXCAPS: A Capsule Network-based Mixture of Experts for Lung Nodule Malignancy Prediction,” *Pattern Recognition*, vol 116, 107942, 2021.

[J5] **P. Afshar**, Sh. Heidarian, F. Naderkhani, A. Oikonomou, K. N. Plataniotis, A. Mohammadi, "COVID-CAPS: A capsule network-based framework for identification of COVID-19 cases from X-ray images," *Pattern Recognition Letters*, vol. 138, pp. 638-643, 2020.

[J4] **P. Afshar**, A. Mohammadi, P.N. Tyrrell, P. Cheung, A. Sigiuk, K. N Plataniotis, E. Nguyen, A. Oikonomou, "DRTOP: Deep learning-based Radiomics for the Time-toevent Outcome Prediction in lung cancer," *Nature Scientific Reports*, vol. 10, 2020.

[J3] **P. Afshar**, A. Oikonomou, F. Naderkhani, P.N. Tyrrell, K. Farahani, K. N Plataniotis, A. Mohammadi, "3D-MCN: A 3D Multi-scale Capsule Network for Lung Nodule Malignancy Prediction," *Nature Scientific Reports*, vol. 10, 2020.

[J2] **P. Afshar**, K. N Plataniotis, A. Mohammadi, "BayesCap: A Bayesian Approach to Brain Tumor Classification Using Capsule Networks," *IEEE Signal Processing Letters*, vol. 27, pp. 12024-2028, 2020.

[J1] **P. Afshar**, A. Mohammadi, K. N. Plataniotis, A. Oikonomou, H. Benali, "From Handcrafted to Deep Learning-based Cancer Radiomics: Challenges and Opportunities", *IEEE Signal Processing Magazine*, vol. 36, no. 4, pp. 132-160, 2019.

Conference Publication

[C7] **P. Afshar**, Sh. Heidarian, F. Naderkhani, M. J. Rafiee, A. Oikonomou, K. N. Plataniotis, A. Mohammadi, "Hybrid Deep Learning Model for Diagnosis of COVID-19 using CT Scans and Clinical/Demographic Data," Accepted in *IEEE International Conference on Image Processing (ICIP)*, 2021.

[C6] **P. Afshar**, K. N Plataniotis, A. Mohammadi, "BoostCaps: A Boosted Capsule Network for Brain Tumor Classification," *42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 11075-1079, 2020.

[C5] **P. Afshar**, A. Oikonomou, K. N Plataniotis, A. Mohammadi, "MDR-SURV: a Multiscale Deep learning-based Radiomics for SURVival prediction in pulmonary malignancies," *IEEE International Conference on Acoustics, Speech and*

Signal Processing (ICASSP), pp. 2013-2017, 2020.

[C4] **P. Afshar**, K. N Plataniotis, A. Mohammadi, “Capsule Networks for Brain Tumor Classification Based on MRI Images and Coarse Tumor Boundaries,” *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1368-1372, 2019.

[C3] **P. Afshar**, K. N Plataniotis, A. Mohammadi, “Capsule Networks' Interpretability for Brain Tumor Classification Via Radiomics Analyses,” *2019 IEEE International Conference on Image Processing (ICIP)*, pp. 13816-3820, 2019.

[C2] **P. Afshar**, A. Mohammadi, K. N. Plataniotis, “Brain tumor type classification via capsule networks”, *IEEE International Conference on Image Processing (ICIP)*, pp. 3129-3133, 2018.

[C1] **P. Afshar**, A. Shahroudnejad, A. Mohammadi, K. N. Plataniotis, “CARISI: Convolutional Autoencoder-Based Inter-Slice Interpolation of Brain Tumor Volumetric Images,” *IEEE International Conference on Image Processing (ICIP)*, pp. 1458-1462, 2018.

Chapter 2

Literature Review

This chapter provides an overview of the radiomics literature [1] mainly focusing on the two categories of Hand-Crafted Radiomics (HCR) and Deep Learning-based Radiomics (DLR).

2.1 State-of-the-Art in Hand-Crafted Radiomics

Studies on hand-crafted radiomics features [16, 17, 19], typically, consist of the following key steps:

1. ***Pre-processing***, introduced to reduce noise and artifacts from the original data and typically includes image smoothing and image enhancement techniques.

2. ***Segmentation***, which is a critical step within the HCR workflow, as typically HCR features are extracted from segmented sections and many tissues do not have distinct boundaries [18]. Although manual delineation is the conventional (standard) clinical approach, it is time consuming and extensively sensitive to inter-observer variability [17], resulting in a quest to develop advanced (semi) automated segmentation solutions of high accuracy that can also generate reproducible boundaries.

   Automatic and semi-automatic segmentation techniques can be either conventional, meaning that pre-defined features are used to classify image pixels/voxels, or deep learning-based, referring to the use of a deep network to segment the image. Conventional techniques can, themselves, lie within three categories of intensity-based [34], model-based, and machine learning methods. In the former category, intensity is used as the main distinguishing feature of the pixels, while in the model-based approaches, the aim is to improve an initial contour, by optimizing an energy function. In machine learning methods, however, a set of features, including intensity and gradient, are extracted from the pixels. These features are then used as the inputs to a machine learning model, such as a Support Vector Machine (SVM), to classify the pixels. Nevertheless, conventional techniques are subject to several shortcomings. For instance, the intensity of the ROI can, sometimes, be similar to other tissues, and therefore, intensity can not be a good discriminator. Furthermore, the formulation of an energy function, in a model-based segmentation, may involve large number of parameters [35], which makes optimization of the energy function difficult and time-consuming. Deep learning methods, on the other hand, are capable of learning the features that can best distinguish pixels, and can be trained in an end-to-end manner. Deep learning approaches, such as different variations of the U-Net [36], "LungNet" architecture [37], DenseNet [38], and hybrid dilated convolutions (HDC) [39] are currently used more often for medical image segmentation.

3. ***Feature extraction***, which is the main step in radiomics workflow and will be discussed in details in Sub-section 2.1.1.

4. ***Feature reduction***, is another critical step in radiomics as although a large number of quantitative features can be extracted from the available big image datasets, most of the features are highly correlated, irrelevant to the task at hand, and/or contribute to over-fitting of the model. To address these issues, radiomics feature reduction techniques are discussed in Sub-section 2.1.2.

5. ***Statistical analysis***, which refers to utilizing the extracted radiomics features in a specific application. We will further elaborate on such radiomics-based statistical analysis in Sub-section 2.1.3.

2.1.1 Radiomics Feature Extraction

During the feature extraction step within radiomics workflow, different types of features are extracted that can be generally classified into three main categories: (1) First

Table 2.1: Different categories of HCR features commonly used within the context of radiomics.

Category	Description	Sub-category
First Order Radiomics	Concerned with the distribution of pixel intensities and use of elementary metrics to compute geometrical features.	
• Shape Features	Quantify the geometric shape of region or volume of interest [17]	Size of the Region of Interest (ROI); Sphericity; Compactness; Total volume; Surface area, Diameter, flatness and; Surface-to-volume ratio [17, 29].
• Intensity Features	Derived from a single histogram generated from the 2D region or the whole 3D volume [17].	Intensity Mean; Intensity Standard Deviation; Intensity Median; Minimum of Intensity; Maximum of Intensity; Mean of Positive Intensities; Uniformity; Kurtosis; Skewness; Entropy; Normalized Entropy; Difference of Entropy; Sum of Entropy, and; Range [17, 29].
Second Order Radiomics (Texture Features)	Concerned with texture features and relations between pixels to model intra-tumor heterogeneity. Texture features are generated from different descriptive matrices [17].	
• Gray Level Co-occurrence (GLCM)	GLCM [29] is a matrix that presents the number of times that two intensity levels have occurred in two pixels with specific distance.	Contrast; Energy; Correlation; Homogeneity;Variance; Inverse Difference Moment; Sum of Average; Sum of Variance; Difference of Variance; Information Measure of Correlation; Autocorrelation; Dissimilarity; Cluster Shade; Cluster Prominence; Cluster Tendency, and; Maximum Probability.
• Gray Level Run-Length (GLRLM)	GLRLM [40] is a matrix that presents the length of consecutive pixels having the same intensity.	Short run emphasis; Long run emphasis; Gray Level Non-Uniformity; Run length non-uniformity; Run percentage; Low gray level run emphasis, and; High gray level run emphasis [17].
• Neighborhood Gray Tone Difference Matrix (NGTDM)	NGTDM [29] is concerned with the intensities of neighboring pixels instead of the pixel itself.	Coarseness; Contrast; Busyness; Complexity Texture Strength.
• Grey-Level Zone Length Matrix (GLZLM)	GLZLM [23] considers the size of homogeneous zones in every dimension.	Zone Percentage; Short-Zone Emphasis; Long-Zone Emphasis; Gray-Level Non-Uniformity for zone; Zone Length Non-Uniformity.
Higher Order Radiomics	Use of filters to extract patterns from images.	Wavelets; Fourier features [29]; Minkowski functionals; Fractal Analysis [18], and; Laplacian of Gaussian (LoG) [23].

order (intensity-based and shape-based features) [24]; (2) Second order (texture-based features) [24], and; (3) Higher order features [18]. Table 2.1 provides a summary of different potential features. It is worth mentioning that HCR features are not limited to this list and can exceed hundreds of features (e.g., in Reference [20] 400 HCR features are initially extracted before going through a feature reduction process). Below, we further investigate the most commonly used categories of hand-crafted features:

1. Intensity-based Features: Intensity-based methods convert the multi-dimensional ROI into a single histogram (describing the distribution of pixel intensities), from which simple and basic features (e.g., energy, entropy, kurtosis, and skewness) are derived. Intensity features allow us to investigate properties of the histogram such as sharpness, dispersion, and asymmetry. These features are, however, the most sensitive ones to image acquisition parameters such as slice thickness [29]. Therefore, designing intensity-based features need special care and pre-processing. Among all intensity features, entropy and uniformity are the most commonly used ones in Radiomics [40]. Generally speaking, entropy measures the degree of randomness within the pixel intensities, and takes its maximum value when all the intensities occur with equal probabilities (complete randomness). Uniformity, on the other hand, estimates the consistency of pixel intensities, and takes its maximum value when all the pixels are of the same value. Although intensity-based features are simple to calculate and have the potential to distinguish several tissues such as benign and malignant tumors [40], they suffer from some drawbacks. First, the selected number of bins can highly influence such features, as too small or too large bins can not resemble the underlying distribution correctly, and as such these features are not always reliable representatives. Besides, optimizing the number of histogram bins can also be problematic, because it leads to different number of bins for different ROIs, and makes it difficult to compare the results of various studies.

2. Shape-based Features: Shape-based features describe the geometry of the ROI and are useful in the sense that they have high distinguishing ability for problems such as tumor malignancy and treatment response prediction [29]. Although radiologists commonly use shape features (also referred to as "Semantic Features" or "Morphological features"), the aim of Radiomics is to quantify them with computer assistance [18]. These features are extracted from either 2D or 3D structures to investigate different shape and size characteristics of the ROI.

Among different shape-based features, volume, surface, sphericity, compactness, diameter, and flatness are more commonly used in Radiomics. For instance, sphericity measures the degree of roundness of the volume or region of interest. Compactness is itself defined based on sphericity and as such, these two need not to be calculated simultaneously, and one of them will be probably excluded by the feature selection methods, which are targeting feature redundancy.

3. Texture-based Features: Shape-based and intensity-based features fail to provide useful information regarding correlations between different pixels across a given image. In this regard, texture-based features are the most informative ones, specially for problems where tissue heterogeneity plays an important role, because texture-based features can catch the spatial relationships between neighboring pixels [29]. In Radiomics, typically, texture-based features are extracted based on different descriptive matrices, among them gray level co-occurrence matrix (GLCM), gray level run length matrix (GLRLM), neighborhood gray tone difference matrix (NGTDM), and gray level zone length matrix (GLZLM) are the most commonly used ones [40], which are defined below:

- **The GLCM**, models the spatial distribution of pixels' intensities and can be calculated by considering the frequency of the occurrence of all pairs of intensity values. Features extracted from GLCM are the most commonly used textural features in Radiomics [40]. Each GLCM is associated with two predefined parameters θ and d, where $\theta \in \{0°, 45°, 90°, 135°\}$, and d is any integer distance admissible within the image dimensions.

- **The GLRLM**, defines the number of adjacent pixels having the same intensity value, e.g., the (i, j) element of the $GLRLM_\theta$ matrix determines the number of times intensity value i has occurred with run length j, in direction θ.

- **The NGTDM**, which is based on visual characteristics of the image, is a vector whose k^{th} element is defined as the summation of differences between all pixels with intensity value k and the average intensity of their neighborhood (size of which is determined by the user).

- **The GLZLM**, which looks for zones in a matrix. A zone can be defined as the set of connected pixels/voxels sharing the same intensity. The $(i, j)^{th}$ element of the GLZLM corresponds to the number of zones with the intensity i, and the size j.

4. Higher Order Radiomics Features: Higher order features such as Wavelet and Fourier features capture imaging bio-markers in various frequencies [29]. Wavelet features are the mostly used higher order features in Radiomics. Wavelet course and fine coefficients represent texture and gradient features respectively, and is calculated by multiplying the image by a matrix including complex linear or radial "wavelet mother functions". Fourier features can also capture gradient information. Minkowski Functional (MF) is another common higher order feature extractor considering the patterns of pixels with intensities above a predefined threshold.

In brief, the MFs are computed by initially forming a binary version of the ROI through utilization of several thresholds within the minimum and maximum intensity limits. Although the number of utilized thresholds is a free parameter, for better results, it should be identified through a selection mechanism (typically empirical tests are used). Based on the binarized ROI, different MFs such as area and perimeter are computable as follows

$$\begin{aligned} MF_{\text{area}} &= n_s, && (2.1)\\ \text{and} \quad MF_{\text{perimeter}} &= -4n_s + 2n_e, && (2.2)\end{aligned}$$

where n_s and n_e are the total number of white pixels (above the threshold) and edges, respectively.

2.1.2 Radiomics Feature Reduction Techniques

Feature reduction is another critical step in radiomics as although a large number of quantitative features can be extracted from the available image datasets, most of the features are highly correlated, irrelevant to the task at hand, and/or contribute to over-fitting of the model (making it highly sensitive to noise). Feature reduction techniques that are used in radiomics can be classified into supervised and unsupervised categories [24], as summarized in Table 2.2. Supervised approaches, such as filtering and wrapper methods, take the discriminative ability of features into account and favor features that can best distinguish data based on a pre-defined class. Unsupervised methods, on the other hand, aim to reduce feature redundancy and include

Table 2.2: Feature reduction techniques commonly used within the Radiomics literature.

Category	**Description**	**Methods**
Supervised	Considers the relation of features with the class labels and features are selected mostly based on their contribution to distinguish classes.	
• Filtering (Univariate)	Test the relation between the features and the class label one by one.	Fisher score (FSCR); Wilcoxon rank sum test; Gini index (GINI); Mutual information feature selection (MIFS); Minimum redundancy maximum relevance (MRMR), and; Student t-test [40].
• Wrapper (Multivariate)	Considers both relevancy and redundancy.	Greedy forward selection, and Greedy backward elimination.
Unsupervised	Does not consider the class labels and its objective is to remove redundant features.	
• Linear	Features have linear correlations.	Principle Component Analysis (PCA), and; Multidimensional scaling (MDS)
• Nonlinear	Features are not assumed to be lied on a linear space.	Isometric mapping (Isomap), and; Locally linear embedding (LLE).

Principle Component Analysis (PCA), Independent Component Analysis (ICA) and Zero Variance (ZV) [24].

In summery, various objectives can be defined when reducing the feature space in radiomics. The following key characteristics can be defined for feature selection purposes [17, 18]:

- *Reproducibility*: Reproducible features (also referred to as "stable features") are the ones that are more robust to pre-processing and manual annotations. These features will be discussed in Sub-section 2.1.4.

- *Informativeness and Relevancy*, which can be defined as features that are highly associated with the target variable [29]. For instance a χ^2-test, calculates the chi-squared statistic between features and the class variable, and consequently features with low impact on the target are discarded. Another selection approach is a Fisher score test, where features with higher variance are treated as the more informative ones.

- *Redundancy*: Non-redundant features are the ones with small correlation with each other. Feature redundancy is defined as the amount of redundancy present in a particular feature with respect to the set of already selected features.

Below, supervised and unsupervised techniques commonly used in Radiomics are further discussed.

1. Supervised Feature Selection Methodologies: Supervised methods are generally divided into two categories as outlined below:

- *Filtering (Univariate) Methods*: These methods consider the relation between the features and the class label one at a time without considering their redundancy. Among all filtering approaches, Wilcoxon test based method has been shown to be more stable, resulting in more promising predictions in the field of Radiomics [40]. A Wilcoxon test is a nonparametric statistical hypotheses testing technique that is used to determine dependencies of two different feature sets, i.e., whether or not they have the same probability distribution.

- *Wrapper (Multivariate) Methods*: Filtering methods have the drawback of ignoring relations between features which has led to development of wrapper techniques. In contrary to the filtering methods, wrapper methods investigate the combined predictive performance of a subset of features, and the scoring is a weighted sum of both relevancy and redundancy [40]. However, computational difficulties prevent such methods from testing all the possible feature subsets.

  Wrappers methods include greedy forward selection and greedy backward elimination. In a forward feature reduction path, selection begins with an empty set and the correlation with class label is calculated for all features individually. Consequently, the most correlated feature is selected and added to the set. In the next step, the remaining features are added, one by one, to this set to test the performance of the obtained set, and the process continues until no further addition can increase the predictive performance of the set. A backward selection path works in contrary to the forward one, beginning with a set including all the available features, and gradually reduces them until no further reduction improves the performance.

Since supervised methods are based on class labels, they are subject to over-fitting and can not be easily applied to different applications once trained based on a given feature set.

2. Unsupervised Feature Selection Methodologies: Unsupervised approaches try to reduce the feature space dimensionality by removing redundant features (those

Table 2.3: Common analysis methods in Radiomics.

Purpose	Description	Methods
Clustering	Similar patients are grouped together based on a distance metrics.	Hierarchical, Partitional
Classification	Models are trained to distinguish patients based on their associated clinical outcome.	Random Forest (RF); Support Vector Machine (SVM); Neural Network (NN); Generalized linear model (GLM); Naive Bayes (NB); k-nearest neighbor (KNN); Mixture Discriminant Analysis (MDA); Partial Least Squares GLM (PLS), and; Decision Tree (DT).
Time-related analysis	The survival time or the probability of survival is calculated based on the available set of data from previous patients.	Kaplan-Meier survival analysis; Cox proportional hazards regression model [19], and; Log-Rank Test.

who are correlated and do not provide any additional information). Although these methods are not prone to over-fitting, they are not guaranteed to result in the optimum feature space. Unsupervised techniques can be divided into linear and non-linear methods, where the former assumes that features lie on a linear space.

2.1.3 Radiomics Statistical Analysis

Statistical analysis refers to utilizing the extracted radiomics features in a specific task. Although most statistical methods, initially, treat all the features equally and use the same weights over all predictors, in the area of radiomics, the most successful methods are the ones that use a prior assumption (provided by experts) over the meaning of features [18]. One basic approach to analyze the radiomics features adopted in [20,23] is to cluster the extracted features and look for associations among clusters and clinical outcomes. For instance, patients belonging to one cluster may have similar diagnosis or patterns. Observations show that image bio-markers are associated with clinical outcomes such as tumor malignancy. Hierarchical clustering is most commonly used in radiomics [17]. However, clustering techniques are not basically trained for target forecasting purposes, which necessitates the use of prediction tools that are specially trained on predefined class label. Prediction tools in radiomics are categorized as:

(i) *Classification and Regression Models* that are mostly similar to other multimedia domains, trying to foresee a discrete or continues value. Random Forest

(RF), Support Vector Machine (SVM) and Neural Network (NN) are among the most common regression and classification techniques used to make predictions based on radiomics [24].

(ii) *Survivability analysis*: Also referred to as time-related models, mostly try to predict the survival time associated with patients. These models are also useful when testing the effectiveness of a new treatment.

Table 2.3 presents a summary of different Radiomics analysis techniques. As predictors belonging to the former category are also common in other multi-media applications, they are not covered in this chapter. Survivability analysis (the latter category), however, is more specific to radiomics. This category includes Kaplan-Meier Survival Curve (KMS), Cox Proportional Hazards (regression) Model (PHM), and Log-Rank Test.

1. Kaplan-Meier Survival Curve (KMS): The KMS curve [20, 24] represents a trajectory for measuring the probability of survival $S(t)$ in given points of time t, i.e.,

$$S(t) = \frac{Number\ of\ patients\ survived\ until\ t}{Number\ of\ patients\ at\ the\ beginning}. \tag{2.3}$$

The KMS curve can be calculated for all radiomics features to assess the impact of different features on patients' survival as follows:

1. A desired feature, for which the KMS curve is supposed to be calculated, is selected.

2. Based on the selected feature, one or more thresholds are considered that can partition patients into, e.g., low and high risk cancer subjects. Patients are then grouped based on whether their associated feature lies above or below the threshold.

3. The KMS curve is calculated for all the obtained groups, and the result can be used to compare the survivability among patients with, e.g., low and high risk cancer. For instance, in Reference [20] *high heterogeneity features* are associated with shorter survival time, while *high compactness features* are associated with longer survival.

*2. **Cox Proportional Hazards (Regression) Model (PHM)*** [20], is commonly used in medical areas to predict patient's survival time based on one or more predictors (referred to as covariates) such as radiomics features. The output of the PHM model denoted by $h(t)$ is the risk of dying at a particular time t, which can be calculated as follows

$$h(t) = h_0(t) \times \exp^{\sum_{i=1}^{N_c} b_i x_i} \tag{2.4}$$

where x_i, for $(1 \leq i \leq N_c)$, are predictors (covariates); b_i represent the impacts of predictors, and $h_0(t)$ is called the base-line hazard. The exponent term in Eq. (2.4) is referred to as the "Risk" and is conventionally assumed to be a linear combination of the features (covariates), i.e., Risk $\triangleq \sum_{i=1}^{N_c} b_i x_i$. The Risk coefficients (b_i, for $(1 \leq i \leq N_c)$) are then computed through a training process based on historical data. More realistically, the risk can be modeled as a general non-linear function, i.e., Risk $\triangleq \boldsymbol{f}(\boldsymbol{x})$, with the non-linearity being learned via deep learning architectures, which has not yet been investigated within the radiomics context.

*3. **Log-Rank Test*** [20], which is used for comparing the survival of two samples specially when these two samples have undergone different treatments. This test is a non-parametric hypothesis test assessing whether two survival curves vary significantly. One limitation associated with the Log-Rank test is that the size of the groups can influence the results, therefore, larger number of patients should be included to from equal sized groups.

2.1.4 Radiomics Stability

An important aspect of radiomics is the stability of the extracted features, which quantifies the degree of dependency between features and pre-processing steps. Stability in radiomics is generally evaluated based on either of the following two techniques:

1. **Test-Retest**: In this approach, patients undergo an imaging exam more than once and images are collected separately. Radiomics features are then extracted from all the obtained sets and analyzed. Here, being invariant across different set of images illustrates stability of radiomics features.

2. **Inter-observer reliability**, which is referred to an experiment where multiple

observers are asked to delineate the ROI from the same images, and radiomics features are extracted from all different delineations to test their stability for variation in segmentation [23]. Here, being invariant across different segmentations illustrates stability of radiomics features.

Different stability Criteria are used to find robust features in radiomics as briefly outlined below:

1. ***Intra-class Correlation Coefficient (ICC)***: One approach to measure the stability of radiomics features, which is used for both the aforementioned categories (i.e., test-retest and inter-observer setting) is referred to as the intra-class correlation coefficient (ICC) [20]. The ICC is defined as a metric of the reliability of features, taking values between 0 and 1, where 0 means no reliability and 1 indicates complete reliability. Defining terms BMS and WMS as mean squares (measure of variance) between and within subjects, which are calculated based on a one-way Analysis of variance (ANOVA), for a test-retest setting, the ICC can be estimated as

   $$ICC_{\text{Test-Retest}} = \frac{BMS - WMS}{BMS + (N-1)WMS}, \tag{2.5}$$

   where N is the number of repeated examinations. By defining EMS as residual mean squares from a two-way ANOVA and M as the number of observers, for an inter-observer setting, the ICC can be calculated as

   $$ICC_{\text{Inter-Observer}} = \frac{BMS - EMS}{BMS + (M-1)EMS}. \tag{2.6}$$

2. ***Friedman Test***: The Friedman test, which is specially useful for assessing the stability in an inter-observer setting, is a nonparametric repeated measurement that estimates whether there is a significant difference between the distribution of multiple observations, and has the advantage of not requiring a Gaussian population. Based on this test, the most stable features are the ones with a stability rank of 1 [20].

In [20], it is declared that radiomics features with higher stability have more prognostic performance, therefore, stability analysis can be interpreted as a feature reduction technique. According to Reference [23], Laplacian of Gaussian (LoG), intensity-based,

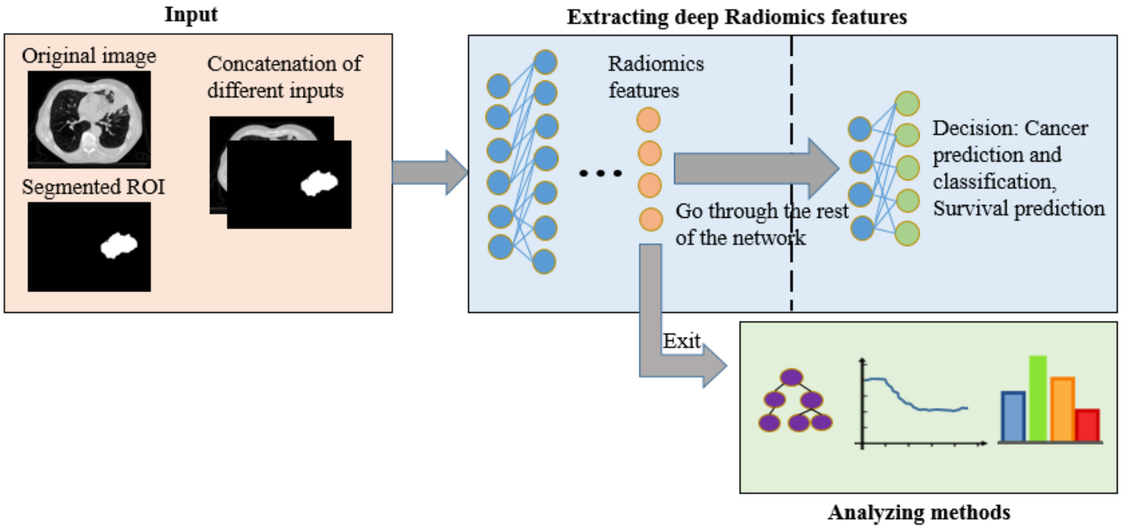

Input
Original image
Concatenation of different inputs
Segmented ROI
Extracting deep Radiomics features
Radiomics features
Go through the rest of the network
Decision: Cancer prediction and classification, Survival prediction
Exit
Analyzing methods

Figure 2.1: Extracting deep radiomics. The input to the network can be the original image, the segmented ROI, or the combination of both. Extracted radiomics features are either utilized through the rest of the network, or an external model is used to make the decision based on radiomics.

and texture features are more stable for lung CT images, while wavelet and shape-based features are sensitive to variation in segmentation.

2.2 State-of-the-Art in Deep Learning-based Radiomics

Deep learning-based radiomics (DLR), sometimes referred to as "Discovery Radiomics" or "Radiomics Sequence Discovery" with "Sequence" referring to features [41], is the process of extracting deep features from medical images based on the specifications of a pre-defined task including but not limited to disease diagnostics; Cancer type prediction, and; Survival Prediction. In brief, the DLR can be extracted via different architectures (stack of linear and non-linear functions), e.g., convolutional neural network (CNN) or an auto-encoder, to find the most relevant features from the input [42]. Fig. 2.1 illustrates the schematic of extracting deep features. The extracted features can then either go through the rest of the deep net for analysis and making decisions or they may exit the network and go through a different analyzer such as an SVM or a Decision Tree (DT). Commonly used deep architectures for radiomics will be discussed in details later in Section 2.2.3.

Benefits of DLR vs. HCR: An important advantage of DLR over its hand-crafted counterpart is that the former does not need any prior knowledge and features can be extracted in a completely automatic fashion with high level features extracted from low level ones [42]. Moreover, deep learning networks can be trained in a simple end-to-end process, and their performance can be improved systematically as they are fed with more training samples [43]. Another key benefit of using DLR instead of HCR is that the input to the deep networks to extract radiomics features, can be the raw image without segmenting the region of interest, which serves the process in two ways:

(i) Eliminating the segmentation step can significantly reduce the computational time and cost by taking the burden of manual delineation off the experts and radiologists, besides, manual annotations are highly observer-dependent, which makes them unreliable sources of information, and;

(ii) Automatic segmentation methods are still highly error prone and inaccurate to be used in a sensitive decision making process.

Furthermore, the input to a deep network can also be the combination of the original and segmented image along with any other pre-processed input such as the gradient image (referred to as "multi-channel" input), all concatenated along the third dimension [21]. The variety of input types can even go further to include images from different angles such as coronal and axial [44].

Generally speaking, studies on DLR can be categorized from several aspects including:

(i) *Input Hierarchy*: The input to the deep net can be the single slices, the whole volume, or even the whole examinations associated with a specific patient. Each of these cases require their own strategy, e.g., in case of processing the whole volume simultaneously, one should think of a way to deal with the inconsistent dimension size, as patients are associated with different number of slices. One common architecture that allows for utilization of inputs with variable sizes, such as various number of slices, is the Recurrent Neural Network (RNN), which will be briefly discussed in Section 2.2.3;

(ii) *Pre-trained and Raw Models*: Depending on the size of the available dataset and also the allocatable time, pre-trained models can be fine-tuned or raw models

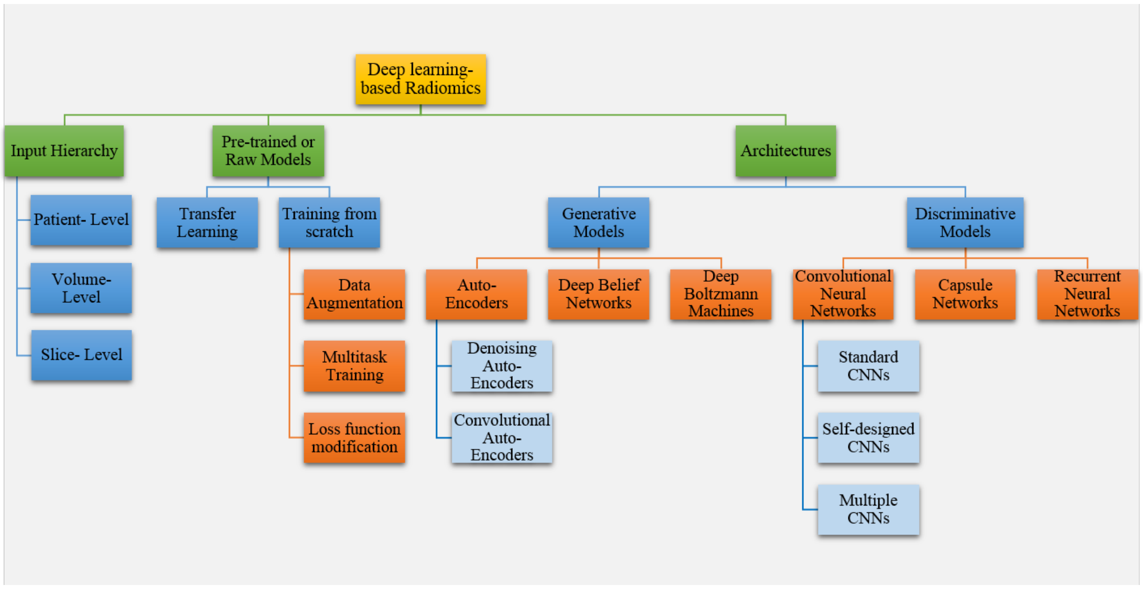


Figure 2.2: Taxonomy of deep learning-based radiomics (DLR).

can be trained from scratch. This will be analyzed more specifically in Section 2.2.2, and;

(iii) *Deep Learning Network Architectures*: Choice of the deep network is the most important decision one should make to extract meaningful and practical DLR, which will be discussed in Section 2.2.3.

2.2.1 Input Hierarchy

As shown in Fig. 2.3, input images for DLR studies can be divided into three main categories: Slice-level; Volume-level, and; Patient-level. Slice-level classification refers to analyzing and classifying image slices independent from each other, however, this approach is not informative enough as we typically need to make decisions based on the labels assigned to the entire Volume of Interest (VOI). Shortcomings of slice-level classification leads to another approach referred to as volume-level classification, where either the slice-level outputs are fused through a voting system, or the entire image slices associated with a volume is used as the input to the classifier. Finally, patient-level classification refers to assigning a label to a patient based on a series of studies (such as CT imaging follow-ups). For example, in Reference [45], patient-level classification is explored with the goal of estimating the probability of lung tumor

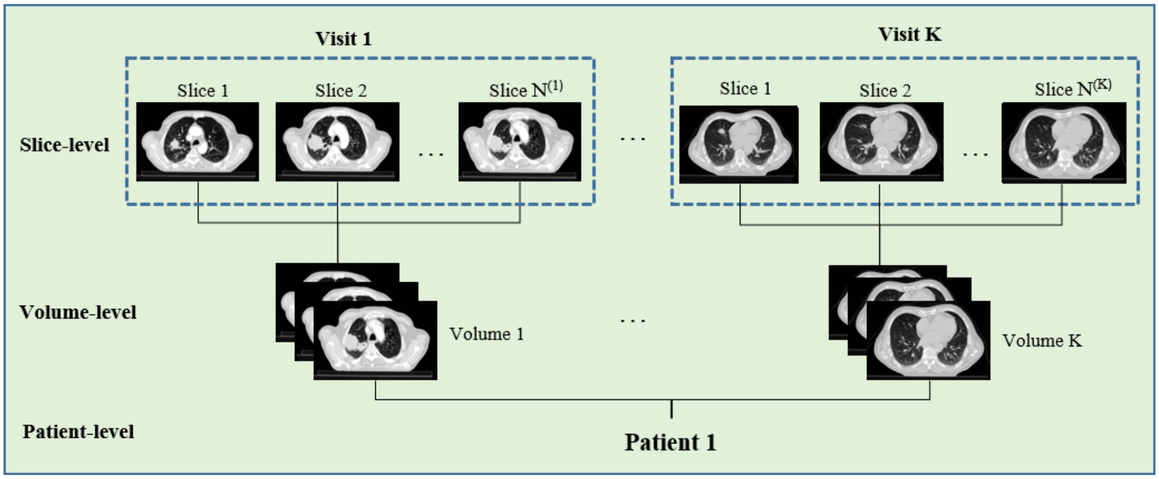


Figure 2.3: Input hierarchy for one patient. In the top row the slice-level input is shown where the patients went through K examination visits during each of which $N^{(i)}$, for $(1 \leq i \leq K)$, number of slices is captured. The second row shows the volume-level where all slices associated with one visit is provided simultaneously as the input to the network. Finally, the third row shows the Patient-level analysis, where a single input consisting of all the volumes is provided.

malignancy based on a set of CT studies. To achieve this goal, initially, a simple three layer CNN is trained to extract DLR from tumor patches associated with individual CT series (volume-level classification) with the objective of minimizing the difference between the predicted malignancy rate and the actual rate. Then, by adopting a previously trained CNN, the malignancy rate is calculated for all the series belonging to the patient and the final decision is made by selecting the maximum malignancy rate. In other words, a patient is diagnosed with malignant lung cancer if at least one of the predicted rates is above a pre-determined rate for malignancy.

2.2.2 Pre-trained or Raw Models

Similar to the other medical areas, the DLR can be extracted based on either of the following two approaches:

Training from scratch: Training a deep network from scratch for extracting DLR has the advantage of having a network completely adjusted to the specific problem at hand. However, performance of training from scratch could be limited due to couple of key issues, i.e., over-fitting and class imbalance. Adhering to patients' privacy and need for experts to provide ground truth typically limits the amount of medical datasets available for extracting DLR resulting in over-fitting of the deep nets. The second issue is the problem of class imbalance, i.e., unequal number of positive and negative classes. This happens as number of patients diagnosed with abnormalities is commonly less that the amount of data available from healthy subjects. More specifically, class imbalance in medical areas is due to the fact that typically number of positive labels is less than the number of negative ones, making the classifier biased toward the negative class, which is more harmful than the other way around because, for instance, classifying a cancerous patient (positive label) as healthy (negative label) has worse consequences than classifying a healthy patient as cancerous [22]. The following strategies can be adopted to address these two issues:

(i) ***Data Augmentation***, where different spatial deformations (such as rotation [46]) are applied to the existing data in order to generate new samples for training purposes. Sub-patch Expansion [21] is another form of augmentation commonly adopted in Radiomics to handle the inadequate data situation via extracting several random fixed-sized sub-patches from the original images.

(iii) ***Multitask training*** is another method introduced to handle class imbalance and inadequate data [47], which is achieved by decreasing the number of free parameters and consequently the risk of over-fitting. For instance, this approach is adopted in [48] for spinal abnormality classification based on MRIs through training a multitask CNN. Multitask in this context refers to performing different classification tasks simultaneously via the same unified network (e.g., the network tries to classify disk grading and disk narrowing at the same time). The loss function is defined as the weighted summation of all the losses associated with different tasks. One important decision to make in multitask learning is the point that branching begins, e.g., in Reference [48], a unified CNN is trained, where all Convolutional layers are shared for performing different tasks and tasks are separated from the point that fully connected layers begin.

(iv) ***Loss function modification***: Another common approach specific to handling class imbalance for DLR extraction is to modify the loss functions by giving more weight to the minority class [48].

Transfer Learning via a Pre-trained Network: A different solution to class imbalance and inadequate training data is “transfer learning" followed by “fine tuning" [21, 47, 49]. The transfer learning phase refers to training the deep net using a natural image data set, and then in the fine tuning phase, the trained network will be re-trained using the desired medical dataset. This strategy is adopted in Reference [49], where a pre-trained CNN is used for breast cancer classification based on mammographic images. The pre-trained CNN used is an Alexnet which is too complicated and prone to over-fitting for small datasets. Therefore, this network is first pre-trained using ImageNet database which consists of more than one million natural images. Pre-trained CNN based on ImageNet is also adopted in [50] for lung cancer survival prediction.

2.2.3 Deep Learning Architectures in Radiomics

Radiomics features can be extracted through both discriminative and/or generative deep learning networks. As is evident from its name, discriminative deep models try to extract features that make the classes (e.g., normal or cancerous) distinguishable, and thus these models can directly classify instances from the extracted features. On the other hand, generative models are unsupervised, meaning that they are trained without considering the class labels. Generally, the goal of these models is to learn the data distribution in a way that enables them to generate new data from the same distribution. In other words, generative models can extract the natural and representative features of the data, which can then be used as inputs to a classifier. Furthermore, in the field of radiomics, it is common [43] to train a generative model and use the learned weights as initial weights of a discriminative model. Below, an introduction to widely used discriminative and generative deep models in radiomics is provided.

1. Discriminative Models: Deep discriminative models try to extract features capable of distinguishing class labels, and the objective is to minimize the prediction error. Below, we will review the convolutional neural networks (CNNs) and Recurrent Neural Networks (RNNs), which are the most popular discriminative architectures in radiomics. Later, we will introduce a recently designed deep architecture referred to as the capsule network (CapsNet) [51] and explain how this new architecture can contribute to the radiomics.

1.1. Convolutional Neural Networks (CNNs): CNN is a stack of layers performing convolutional filtering combined with nonlinear activation functions and pooling layers [47]. The fact that CNNs have recently resulted in promising outcomes have made them the mostly used architecture in medical areas including radiomics. CNNs are more practical in the sense that shared weights are utilized over the entire input, which reduces the number of trainable parameters. Unlike extracting hand-crafted features, kernels used in convolutional layers are not pre-determined and are automatically learned through the training process. This property makes CNNs suitable methods for extracting DLR features as they are flexible and can be applied without requiring a prior knowledge. In [52], it has been shown that the DLR extracted from a CNN can visually distinguish benign and malignant lung tumors when projected into a 2D space, while the original pixel values completely fail to provide such distinction.

When adopting CNNs in the field of Radiomics, output of the fully connected layers is typically treated as DLR features. These features are then either used within the original CNN to provide the desired (classification and/or regression) output such as cancer type, or exist the network to be provided as the input to the rest of the Radiomics pipeline. As an example, in Reference [50] DLR are extracted from the layer just before the classification (SoftMax) layer of the CNN with the goal of lung cancer survival prediction. These features are referred to as the "preReLU" and "postReLU" features as they are extracted both before and after applying the ReLU activation function. The DLR features are then used as inputs to four classifiers (i.e., Naive Bayes, Nearest Neighbor, Decision tree and Random Forest) after going through a feature selection algorithm.

The CNN architectures used in Radiomics can be divided into three main categories: (i) Standard architectures; (ii) Self-designed architectures, and; (iii) Multiple CNNs. Below, we describe each of these categories with examples from Radiomics:

1.1.1. ***Standard CNN Architectures***: As the name suggests, standard architectures are those that have been previously designed to solve a specific problem, and due to their success are now being adopted in the Radiomics. Two of such architectures that have been used in Radiomics are LeNet and AlexNet. The LeNet is one of the simplest CNN architectures, having a total of 7 layers, that has been used in Radiomics. However, researchers have some times modified this network to achieve higher performance, e.g., the CNN used in Reference [21] is a LeNet architecture with a total of 9 layers including 3 Convolutional layers, 3 pooling layers and one fully connected layer followed by the classification layer to classify lung tumors as either benign or malignant.

Another commonly used standard architecture in Radiomics is the 11 layers CNN called Alexnet, which has been adopted in [49] to extract DLR features from breast mammographic images. Features are extracted from all 11 layers and used as inputs to 11 support vector machine (SVM) with the goal of classifying breast tumors as either benign or malignant. Since it is not obvious which set (output of which of the 11 underlying layers) of DLR features are more practical, these SVMs are compared and the one with the largest area under the curve is chosen for predictive analysis of breast cancer. The results of [49] concluded that the features extracted from the 9th layer (a fully connected layer before the last fully connected layer and the classification layer) are the best predictors of breast cancer and they are of lower dimension compared to previous ones, which reduces the computational cost. In other words and in contrary to [50], the output of the last Convolutional layer, right before the fully connected layer, is selected as the DLR features.

Although AlexNet is a powerful network, it has too many parameters for small datasets and is, therefore, prone to over-fitting. As a result, Reference [53] has used an Alexnet with number of layers reduced to 5 in order to avoid the over-fitting problem. The input to this network is a combination of CT and PET images, each having 3 channels: One slice corresponding to the center of the lung nodule, specified by an expert, and the two immediate neighbors. The goal of this article is to classify lung tumors as benign or malignant, and although it has been shown that the adopted CNN does not result in significantly higher accuracy than classical methods (HCR), it is more convenient as it does not require the segmented ROI.

Inception network [54,55] is another CNN adopted in Radiomics. This network involves parallel convolutions with different kernel sizes, and poolings within the same layer, with the overall aim of allowing the network to learn the best weights and select the most useful features. The Inception CNN is used in [56], for the detection of diabetic retinopathy. This paper is the first work on deep learning-based detection of diabetic retinopathy that has been approved by the Food and Drug Administration (FDA).

1.1.2. ***Self-designed CNNs***: As opposed to researchers that have used standard CNNs with or without modifications, some have designed their own architectures based on the specification of the Radiomics problem at hand. For example, Reference [46] has used a CNN with three Convolutional layers to extract DLR features, and although the CNN itself is trained to use these features for classifying benign and malignant tumors, they are used as inputs to a binary decision tree.

In a similar fashion, Reference [57] has used a CNN with 6 Convolutional layers and one fully connected layer for DLR extraction in the problem of brain tumor classification. The designed network, however, is different from previously mentioned articles as it is developed for tumor segmentation, and *features are extracted from the last Convolutional layer since they are more robust to shifting and scaling of the input.* In other words, the CNN was designed for segmentation and once trained, the output of the last Convolutional layer is used as the DLR features. The claim here is that the quality of extracted features depends on the accuracy of segmentation, and when segmentation is precise the quality of Radiomics features is guaranteed.

Due to the high importance of the segmentation, more advanced and efficient CNN architectures have been developed, one of which is the Fully Convolutional Neural Network (FCNN) [58]. In an FCNN, fully connected layers are rewritten as convolutional layers, having the advantage of not requiring fixed-sized inputs. This network is also extended to 3D image segmentation, to segment multiple targets at once. To decrease the false positive rate, FCNN is further combined with graphical models such as Markov Random Fields (MRFs) and Conditional Random Fields (CRFs). Finally, to improve the resolution of the output, U-Nets [36] are proposed, which include up-convolutions to increase the image size, and skip-connections to recover spatial information.

Lung cancer detection using CNNs is also investigated in [59], with the difference that the input to the network is not only the original image but also the nodule-enhanced and vessel-enhanced images, stating that providing the network with

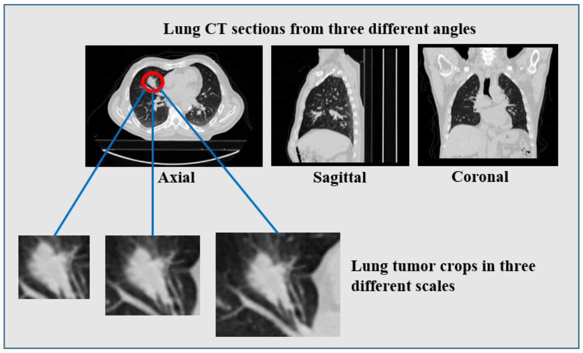

Lung CT sections from three different angles
Axial
Sagittal
Coronal
Lung tumor crops in three different scales

Figure 2.4: Different angles of lung CT scan along with tumor crops in three different scales.

more information on tumor and vessels reduces the risk of misplacing these two by the network. The main focus here is to reduce the false positive rate while keeping the sensitivity high, therefore, a significant number of nodule candidates are selected at the beginning. Use of CNNs is further investigated in [60], where a 7 layer architecture is fed with down-sampled volumetric CT images along with their segmentation masks for longevity prediction. In [61] an architecture called XmasNet is provided that can maximize the accuracy of prostate cancer diagnosis. This network consists of 4 Convolutional layers, 2 fully connected layers, 2 pooling layers and one SoftMax layer for cancer prediction. The inputs to this network are 3D MRI images.

In summary, self-designed CNNs are developed by varying the depth of the network (number of the Convolutional and non-Convolutional layers); the order the layers are cascaded one after another; the type of the input to the network (e.g., single channel or different form of multi-channel), and/or; the layer whose output is treated as the DLR features.

1.1.3. ***Multiple CNNs***: Beside using single standard or self-designed CNNs, some researchers have proposed to use multiple networks, which has the advantage of benefiting from multiple inputs, having various modalities, scales and angles as shown in Fig. 2.4, or different architectures with different properties.

“Scale” is a significant factor to consider when designing the input structure. For example to distinguish tumors from vessels, a large enough region should be included in the input patch, while to differentiate between solid and non-slid tumors, the nodule region should be the main core of the patch. Having this in mind, Reference [44] has designed a CNN architecture for lung tumor classification, where inputs are patches not only from different angles (sagittal, coronal, and axial) but also in different scales. Following a similar path, Reference [62] has also designed a multiple CNN architecture, where each CNN takes a lung tumor patch at a specific scale (illustrated in Fig. 2.4) as input and generates the associated DLR features. Features extracted from all the CNNs are then concatenated and used for lung tumor malignancy prediction through a conventional classifier (SVM). The idea here is that segmenting the tumor regions is not always feasible. Furthermore using a tumor patch provides information on not only the tumor itself but also the surrounding tissues, and since tumor sizes can vary significantly among patients, using multi-scale patches instead of the single ones will improve the overall performance of the extracted DLR features. An interesting property of such multiple CNN architecture is that since the constituent CNNs share parameters, training can be performed in a reasonable time. Another benefit of using a multiple CNN architecture is that the network becomes robust to addition of small noise to the input.

Similar to the work in [62], Reference [63] has designed a CNN called “Multi-view CNN", which uses 7 patches at different scales as inputs, with the difference that these patches are resized to have the same dimension, and therefore, a single CNN can be used instead of multiple CNNs. This work has also extended the binary lung tumor classification to a ternary classification to classify lung tumors as benign, primary malignant, and metastatic malignant. Furthermore, this article has adopted another validation approach called “separability" besides the common terms such as accuracy and AUC (area under curve). Separability refers to the extend that different classes are distinguishable based on the learned features, and according to the aforementioned article, the proposed multi-view CNN has a higher Separability compared to a single scale CNN. In addition to that, as the layers go deeper, features with higher separability are learned.

The idea of using multi-scale image patches is further expanded in Reference [52] through designing a novel CNN architecture called “Multi-crop CNN", where instead of taking inputs in various scales, multi-scale features are extracted through parallel pooling layers, one of which applies pooling to a cropped version of the input from the previous layer. Features from multiple pooling layers are then concatenated and fed to the next layer. 3D lung CT images are inputs to this network, and since multiple CNNs are replaced with one single CNN, the training can be performed in a more time effective manner. Beside forecasting the lung tumor malignancy, this work has also predicted other attributes associated with tumor such as diameter, by replacing the final SoftMax layer with a regression one. It is worth mentioning that this network is not performing all the assigned tasks simultaneously. Instead they are performed one after another, which distinguishes this network from a multitask training one discussed in section 2.2.2.

Radiomics through multiple CNNs is further explored recently in [64] for Alzheimer's disease diagnosis using MRI, where in the first stage several landmarks are detected based on the comparison between normal and abnormal brains. These landmarks are then used to extract patches (separately around each individual landmark), and consequently each CNN is trained taking patches corresponding to a specific landmark position as input. Final decision is made based on a majority voting among all the CNNs. Here, the idea behind using a multiple architecture is the fact that detecting Alzheimer's disease requires the examination of different regions of the brain.

In summary, multiple CNNs methods developed for DLR feature extraction are designed by either fusing the outputs of several CNNs which are trained based on a specific input, or multi-path layers are embedded within a single network to modify the output from previous layers differently.

One challenge shared among all the aforementioned CNN architectures is that they do not take the spatial information between objects into account. As an example, they may fail to consider the location of abnormality within the tissue as an indicator of its type. The newly proposed deep architecture called CapsNets, described next, is introduced to overcome this drawback.

1.2. Capsule Networks: Although CNNs are the state of the art in many medical and non-medical classification problems, they are subjected to several drawbacks

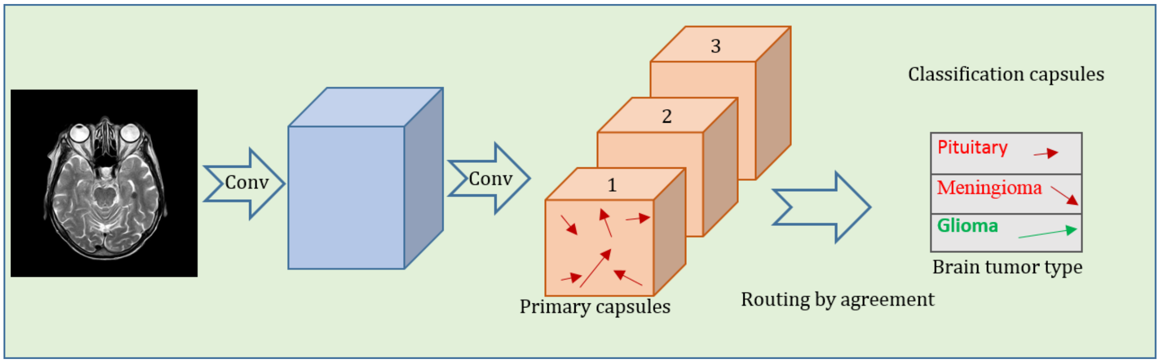


Figure 2.5: Capsule network architecture.

including their low explainability and their negligence in preserving the spatial relationships between elements of the image leading to miss-classification. Besides, CNNs have low robustness to some types of transformation. Loss of spatial relation information, which is associated with the pooling layers, is resolved by the newly proposed capsule networks (CapsNets) [51] consisting of both convolutional and capsule layers that can handle more types of transformation. These deep architectures have the ability to consider the relationships between the location of objects and tolerate more types of transformation, through their routing by agreement process, which dictates that an object will not be classified as a specific category unless the lower level elements of this object agree on the existence of that category. Another important property of CapsNets is that they can handle smaller datasets, which is typically the case in most medical areas. Here we explain the architecture of capsule networks, as illustrated in Fig. 2.5, and their routing by agreement process.

Capsule networks are constructed based on capsules, as their main building blocks. A capsule being represented by a vector consists of several neurons representing, collectively, a specific object at a specific location. While neurons capture the instantiation parameters of the object, the length of a capsule determines the existence probability of that object. The most important property of a capsule network, distinguishing it from CNNs, is its routing by agreement process. Generally speaking, each Capsule i, having the instantiation parameter vector $\boldsymbol{u}_i$, in a lower layer tries to predict the output of the capsules in the next layer, through a trainable weight matrix $\boldsymbol{W}_{ij}$ given by

$$\hat{\boldsymbol{u}}_{j|i} = \boldsymbol{W}_{ij}\boldsymbol{u}_i, \tag{2.7}$$

where $\hat{\boldsymbol{u}}_{j|i}$ denotes the prediction for parent Capsule j. Through the routing by agreement process, the predictions are evaluated in terms of their similarity to the actual outputs. More weight is then given to the successful predictions, before calculating the final output $\boldsymbol{s}_j$ for the capsule j, as follows

$$\begin{aligned} a_{ij} &= \boldsymbol{s}_j.\hat{\boldsymbol{u}}_{j|i}, && (2.8)\\ b_{ij} &= b_{ij} + a_{ij}, && (2.9)\\ c_{ij} &= \frac{\exp(b_{ij})}{\sum_k \exp(b_{ik})}, && (2.10)\\ \text{and } \boldsymbol{s}_j &= \sum_i c_{ij}\hat{\boldsymbol{u}}_{j|i}, && (2.11) \end{aligned}$$

where a_{ij} shows the agreement between actual output $\boldsymbol{s}_j$ and its prediction $\hat{\boldsymbol{u}}_{j|i}$, and c_{ij} denotes the score assigned to the prediction based on the obtained agreement. Routing by agreement, as defined through Eqs. (2.8)-(2.11), is an iterative process, with $\boldsymbol{s}_j$ initially defined as an average over all the predictions. The routing by agreement process, summarized in Fig. 2.6, enables capsule the network to recognize spatial information between image instances. Finally, the margin loss function l_j is computed as

$$l_j = T_j \max(0, m^+ - ||\boldsymbol{s}_j||)^2 + \lambda(1 - T_j)\max(0, ||\boldsymbol{s}_j|| - m^-)^2. \qquad (2.12)$$

Term T_j is 1 whenever the class j is present, and is 0 otherwise. Term m^+, m^- and λ are hyper parameters to be indicated before the learning process. The total loss is the sum over the losses of all output capsules. The original Capsule network has also a set of fully connected layers, referred to as the decoder part, that takes the final instantiation parameters of the true classes as inputs, and try to reconstruct the original image, with the aim of forcing the network to capture real representative features. The decoder loss is defined as a simple squared error and contributes to the final error with a smaller weight, compared to the loss of the capsules. This is done to avoid distracting the network from its main target, which is classifying the objects. This completes a brief introduction to CNNs and CapsNets. Next, we present the proposed framework for tumor classification.

1.3. Recurrent Neural Networks: Most of the deep network architectures need

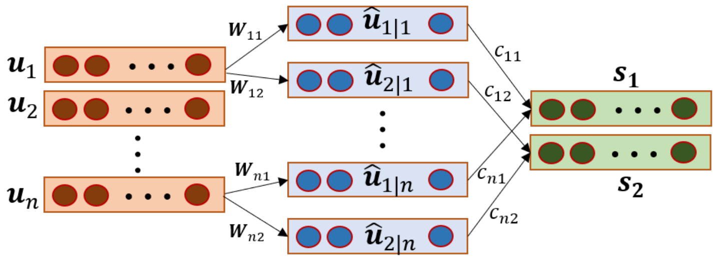


Figure 2.6: Routing by agreement. For the sake of simplicity number of output capsules is set to two.

fixed-sized inputs, which makes them ineffective for Radiomics analysis of volumetric images (volume-level classification), i.e., when the whole volume is needed to be processed at once (such as tumor classification based on the 3D volume). In these scenarios, the RNNs can be adopted as they are capable of processing sequential data such as CT or MR slices, and they take both the present image slice and result of processing the previous ones as inputs. RNNs are also useful to monitor the medical images resulted from follow-up examinations (patient-level classification).

Since RNNs are associated with the vanishing gradient problem, a new type called long-short-term-memory (LSTM) is proposed which has the ability to decide what to store and what to forget. Although it seems that RNNs and LSTM are computationally more expensive than other architectures, their training time and cost is greatly reduced by using the same weights over the whole network [47]. Use of LSTMs is explored in Reference [65] for prostate cancer benign and malignant classification based on sequences of ultrasound images, where it has been shown that the predictive accuracy of this sequential classification is higher than making decision based on independent single images.

2. Generative Models: The objective of most of the deep generative models is to learn abstract yet rich features from the data distribution in order to generate new samples from the same distribution. What makes these models practical in radiomics is the fact that the learned features are probably the best descriptors of the data, and thus have the potential to serve as radiomics features and contribute to a consequent tasks such as tumor classification. Auto-encoder networks, deep belief networks, and deep Boltzmann machines are among the deep generative models that have been utilized in radiomics works as outlined below:

2.1. ***Auto-Encoder Networks***: An auto-encoder network consists of two main components: An encoder which takes as input N_s medical images denoted by $f^{(i)}$, for $(1 \leq i \leq N_s)$, and converts each into a latent space $\phi(\boldsymbol{W} f^{(i)} + \boldsymbol{b})$, i.e., Radiomics features. The second component, the decoder, takes the latent space and tries to reconstruct the input image with the objective of minimizing the difference between the original input and the reconstructed one $\phi(\boldsymbol{W}^T \phi(\boldsymbol{W} f^{(i)} + \boldsymbol{b}) + \boldsymbol{c})$ [42] given by

$$\min_{\boldsymbol{W},\boldsymbol{b},\boldsymbol{c}} \sum_{i=1}^{N_s} ||\phi\left(\boldsymbol{W}^T \phi\left(\boldsymbol{W} f^{(i)} + \boldsymbol{b}\right) + \boldsymbol{c}\right) - f^{(i)}||, \tag{2.13}$$

where $\phi(\cdot)$ is the network's activation function; $\boldsymbol{W}$ denotes the weight matrix of the network used by both the encoder and the decoder; Term $\boldsymbol{b}$ denotes the encoder's bias vector; $\boldsymbol{c}$ is the decoder's bias vector, and; superscript T denotes the transpose operator. The reason that the encoded variables can be treated as Radiomics features is that they are the most important representatives of the input image that can be used to reproduce it. Although an auto-encoder can be trained completely in an end-to-end manner, to begin training with good initial weights and thus avoid the vanishing gradient problem, one can first train layers one by one, and use the obtained weights as the auto-encoder starting point [47]. Depending on the application, Auto-encoders have several extensions including:

2.1.1. **Denoising Auto-Encoders (DAEs)**: To make auto-encoders capture more robust features of the input, one common strategy is to add some noise to the input. This kind of auto-encoder is called a denoising auto-encoder (DAE) [47]. Reference [66] has adopted DAE for extracting Radiomics features that are fed to an SVM to classify lung tumors as benign or malignant. Reference [21] has also adopted a five layer denoising auto-encoder which takes the corrupted lung images as inputs and tries to recover the original image. In particular, 400 Features extracted by the encoder part of this network are treated as Radiomics to train another neural network for lung cancer classification (identify the type of the tumor such as benign or malignant).

2.1.2. **Convolutional Auto-Encoders (CAEs)**: This type of auto-encoders are specially useful for Radiomics (image type inputs) as the spatial correlations are taken into account. In these networks, nodes share weights in a local neighborhood [47]. A CAE with 5 Convolutional layer is adopted in Reference [22] for lung cancer diagnosis (identify the presence of cancer).

There are two common strategies to leverage Auto-encoders in Radiomics:

- The first and most frequent approach is to directly use the extracted features to train a classifier. For instance, [42] has extracted Radiomics features using a 5 layer auto-encoder, which receives the segmented region of interest as the input. These features go through a binary decision tree in the next step to produce the output which is the classified lung nodule in this case.
- Auto-encoders can also serve as a pre-training stage to make the network extract representative features before trying to perform the actual classification. For instance, Reference [43] has first trained a DAE based on resized (down-sampled images to facilitate training) lung CT patches. In the next stage, a classification layer is added to the network and the whole network is re-trained taking both resized images and the resizing ratio as inputs.

2.2. ***Deep Belief Networks (DBNs)***: DBNs are stack of Restricted Boltzmann Machines (RBMs) on top of each other where the RBM is an unsupervised two layer stochastic neural network that can model probabilistic dependencies with the objective of minimizing the reconstruction error. More importantly RBM is a bipartite graph allowing value propagation in both directions. Although DBNs are composition of RBMs, only the top two layers have undirected relations. DBNs are first trained in a greedy fashion meaning that RBM sub-networks are trained individually followed by a fine-tuning phase [47]. In Reference [21], a DBN consisting of 4 hidden layers is designed with the goal of extracting the DLR from the top layer which has 1600 nodes. This last layer is connected to an external neural network to classify lung nodules. Besides, to have multi-channel input (original image, segmented tumor, and gradient image), these channels are concatenated vector wise before being fed to the network.

2.3. ***Deep Boltzmann Machine (DBMs)***: DBMs are also based on RBMs, but they differ from DBNs in the sense that DBMs include undirected relations between all layers which makes them computationally ineffective, though they are trained in a layer wise manner [47]. Due to the two-way relations, however, RBMs can capture complicated patterns from the data [32]. DBMs are adopted in [32] for Alzheimer's disease diagnosis. In this work, a classification layer is added to the last layer of the DBM allowing to extract not only hierarchical (generative) but also discriminative features.

2.2.4 Explainability of Deep Learning-based Radiomics

Explainability of deep networks refers to revealing an insight of what has made the model to come into a specific decision, helping with not only improving the model by knowing what exactly is going on in the network, but also detecting the failure points of the model. No matter how powerful DLR are, they will not be utilized by physicians, unless they can be interpreted and related to the image landmarks used by the experts. Besides, not even a single mistake is allowed in medical decisions as it may lead to a irreparable loss or injury, and having an explanation of the logic behind the outcome of the deep net is the key to prevent such disasters. This subsection will present an overview on recently developed techniques to increase the explainability of deep Radiomics.

One simple approach to ensure the accuracy of the automatic prediction, is to double-check the results with an expert. For instance, Reference [60], which has used a CNN for longevity prediction using CT images, has reviewed the outcomes with experts leading to the fact that people predicted with longer lives are indeed healthier. However, this approach is time consuming and needs complete supervision and investigation, which is in conflict with the concept of automatizing and personalized treatment, which is the whole point of Radiomics. Therefore, nowadays several criteria are being presented to reduce the time and complexity of explaining deep Radiomics. One of these approaches is "feature visualization" which tries to gain knowledge on the network behavior by visualizing what kinds of features the network is looking for. This technique can be applied to different layers of the model. For example, to visualize the first layer features, the associated filters are applied to the input and the resulting feature maps are presented. However, as the last layer is the most responsible one in the network's output, paying attention to the features learned in this layer is more informative. For instance, Reference [21] has visualized the final weights of a DBN, showing that the network is looking for meaningful features such as curvity. Nevertheless, these features are not as meaningful as they are for simple image recognition tasks as clinicians themselves are sometimes unsure about the distinctive properties of the images.

One other method to provide the user with an explanation on the decision made by a deep architecture is called "sensitivity analysis" referring to generating a heat-map highlighting the image regions responsible for the output [48]. In the heat-map, the brighter areas are the ones that have influenced the prediction. This can be achieved by determining and measuring the effect of changing each individual input pixel on the output. In a CNN this effect can be estimated by determining the weight associated with each input pixel through back propagation. This approach can discover the cause for the prediction [48], however, the drawback of this approach is that not all the detected pixels through the heat-map are necessarily the ones leading to the specific decision, and besides, as the depth and complexity of the deep net increases, it becomes more difficult to measure the contribution of each individual pixel on the output.

A third proposed approach to understand the learned features is to project the high-dimensional feature space from the deep network to a bi-dimensional plane. Reference [44] has adopted this strategy by using t-Distributed Stochastic Neighbor Embedding (t-SNE) algorithm to visualize the features learned by a CNN for lung tumor classification. The resulted plane presents clearly defined clusters of lung tumors, which shows that the networks has successfully learned discriminating features. However, although this method can verify the accuracy of the network, it does not provide information on the exact reason behind making the decision.

The interpretability of meaningless weights is improved in the newly proposed Capsule networks through reconstructing the input image based on the features learned by the network. CapsNet includes a set of fully connected layers that take the final calculated features, based on which the final classification is made, as inputs, and reproduce the original image with the objective of minimizing the difference between the original and the reconstructed image. This objective function is added to the classification loss with a smaller weight not to distract network from extracting discriminative features. If the trained CapsNets is not only of high accuracy but also capable of resembling the input image, it has been successful in extracting representative features. Besides, visualizing these features provides insight on the explainability of the model. Interestingly, CapsNets are equipped with a powerful feature visualization technique through their input reconstruction part, which works as follows:

1. If CapsNet is in the training phase, the feature vector associated with the true class label is selected, otherwise, the one with the higher probability is used.

2. The selected feature vector is tweaked, meaning that small numbers are added to or subtracted from the feature values leading to a slightly changed new feature vector.

3 The new feature vector is fed to the reconstruction part and the input image is reproduced. However this reconstructed image is not supposed to exactly resemble the input image as it is generated using the tweaked features not the actual ones learned by the network.

4. By repeating the process of tweaking and reconstructing process over and over again, one can understand what features are learned by observing the influence of changing them on the generated images.

2.3 Hybrid solutions to Radiomics

To summarize our findings on HCR and DLR features, Table 2.4 provides different comparisons between these two categories from various perspectives. In scenarios where neither of the above two mentioned categories are capable of providing informative Radiomics features with high predictive capacity, one can resort to hybrid strategies. Here, potential hybrid solutions to Radiomics [67] are reviewed from different points of view including combination of Radiomics with other data sources and combination of HCR and DLR features.

Table 2.4: A Comparison between hand-Crafted and Deep Radiomics.

Hand-Crafted Radiomics (HCR)	**Deep Radiomics (DLR)**
Needs a prior knowledge on types of features to extract.	Can learn features on its own and without human intervention.
Features are typically extracted from the segmented ROI.	does not necessarily require a segmented input.
It is generally followed by a feature selection algorithm.	Feature selection is rarely performed.
As features are defined independent from the data, does not require big datasets.	Requires huge datasets, since it has to learn features from the data.
Processing time is not normally significant.	Can have high computational cost depending on the architecture and size of the dataset.
Since features are pre-designed, they are tangible.	The logic behind the features and decisions is still a black box.

2.3.1 Combination of Radiomics and Other Data Sources

Physicians, normally, do not rely on a single input for their diagnosis of diseases and disorders. To come into a conclusive decision, inputs from different sources are compared and combined including Radiomics (image bio-markers from different imaging modalities); Blood bio-markers; Clinical outcomes; Pathology, and; Genomics results [16]. Below, we discuss two different ways to fuse/combine Radiomics with other available resources of information along with the rationales and potentials behind such combinations:

i. ***Extracting Radiomics from Different Imaging Modalities***: As stated previously, Radiomics can be extracted from different imaging modalities each of which can only capture/provide particular information on tissues' properties. For instance, although the CT scan is among the most common and informative imaging modalities allowing to observe the body internal organs, the CT can not provide information on body function and metabolism. This type of information is available through PET scan, which calls for studying the effect of combining

Radiomics extracted from different modalities. For example, in [19] Radiomics features are extracted from both CT and PET images, and the concatenated feature vector is fed to a classifier for lung cancer survival prediction, resulting in a higher accuracy compared to each modality separately.

Combining different imaging modalities is also tested on brain tumor classification in [57]. Since MRI can output different images varying mostly in terms of their contrast and relaxation parameters, these images can be fused to provide complementary information. Based on [57], extracting Radiomics from this combination of MRIs outperforms the single modal classifier for brain tumor classification.

ii. ***Integration of Extracted Radiomics with Other Data Sources***: Radiomics features are combined with other resources only after the extraction process. The best descriptive or predictive models in the field of Radiomics are the ones that utilize not only imaging bio-markers, but also other information such as Genomics patterns and tumor histology [18]. In [20], it is reported that combining Radiomics features with lung cancer staging information, which is obtained based on the tumor location and dispersion, can improve the prognostic performance of Radiomics alone or staging alone. In other words, Radiomics can provide a complementary information for lung cancer prognosis [17]. It is also shown that combining Radiomics with other prognostic markers in head-and-neck cancer leads to a more inclusive decision. Combining Radiomics with clinical data is further investigated in [68] for brain cancer survival prediction. The interesting output of this work is a nomogram based on both Radiomics and clinical risk factors such as age that can be used to visually calculate survival probability. Developing such a nomogram is further investigated in references [69] and [70] for prediction of Hepatitis B virus and lymph node metastasis, respectively.

In brief, the first step to build a Radiomics-based nomogram, is calculating a linear combination of selected Radiomics features based on a logistic regression, which results in a Radiomics score to exploit further for the desired prediction task. Consequently, by training a multivariate logistic regression, Radiomics score is fused with other influential factors to make the final prediction. Fig. 2.7

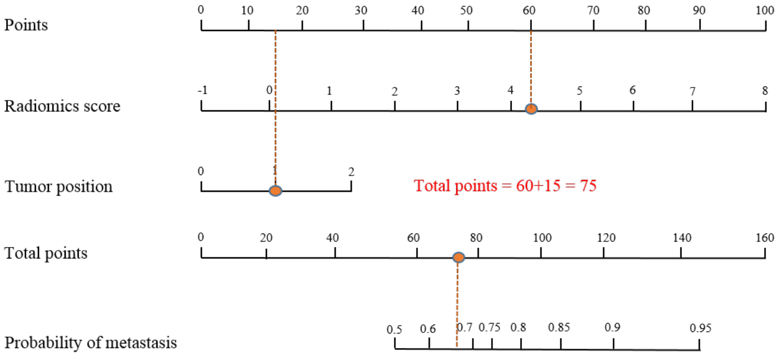


Figure 2.7: Radiomics-based nomogram to predict lymph node metastasis. Tumor position is considered as extra information to assist with making a more reliable prediction.

presents the nomogram introduced in [70] along with an example illustrating how the lymph node metastasis prediction is made.

Reference [67] has adopted a fusion approach based on both Radiomics and Genomics bio-markers for predicting the recurrence risk of lung cancer. In this study, 35 hand-crafted features are extracted from segmented lung CT images and reduced to 8 after a feature selection phase. These features are then used to train a Naive Bayesian network. The same classifier is also trained using two Genomics bio-markers, and the outputs of two classifiers are fused through a simple averaging strategy. Results demonstrate that the combination of classifiers not only leads to higher prediction accuracy compared to individual ones, but also resembles the Kaplan-Meier plot of survival more precisely.

2.3.2 Fusion of HCR with DLR (i.e., Engineered Features Coupled with Deep Features)

As mentioned in Table 2.4, engineered (hand-crafted) and deep Radiomics both have their own advantages and disadvantages. As a result, combining these features has the promise of benefiting from both domains and incorporating different types of features [59] potentially results in significantly improved performance. As shown in Fig. 2.8, the following two categories of data fusion have been used in Radiomics most recently:

1. **Decision-level Fusion**: One common approach to combine HCR with DLR is to

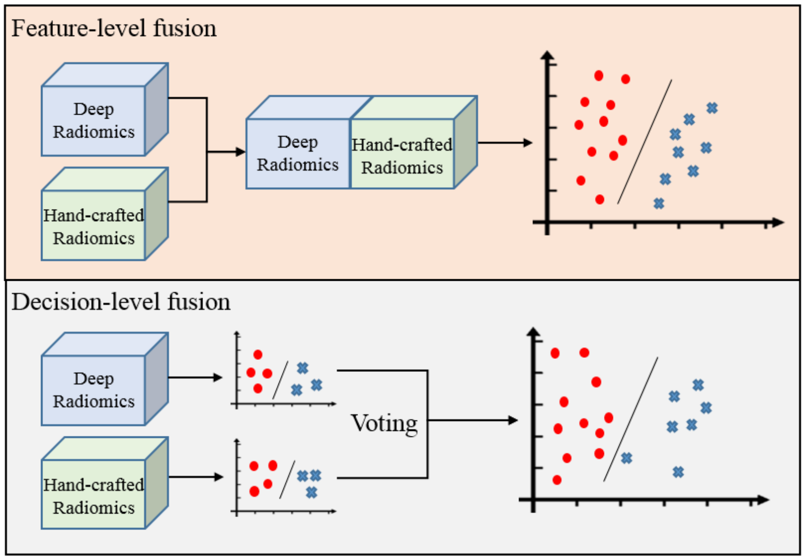


Figure 2.8: Combining deep and hand-crafted Radiomics through feature-level or decision-level fusion.

first use them separately to train separate classifiers and then adopt a kind of voting between the outputs to make the final decision. The voting or fusion approaches in Radiomics include:

(i) ***Soft Voting***, which is combining the probability outputs, for instance, through a simple averaging. Soft voting is adopted in [49], where two individual SVMs are trained on hand-crafted and deep Radiomics features (extracted using a pre-trained CNN), and consequently breast cancer prediction is performed based on averaging the output probabilities. Results of this article shows that the combined features are associated with higher prediction accuracy. Fusion of separately trained classifiers through soft-voting based on deep and hand-crafted Radiomics is also examined in [71] for breast cancer classification, where it has been shown that the combined SVM model outperforms individual classifiers in term of accuracy for mammogram, ultrasound, and MRI images.

(ii) ***Hard Voting***, which is combining outputs, for example, through a majority vote.

(iii) ***Adaptive Voting***, where a weight of importance for each model (HCR and DLR) is learned for example using a separate neural network. In Reference [72], a different kind of voting is adopted for lung cancer classification. This voting is based on the idea that not all the classifiers contribute equally to the final decision, and contribution weights are parameters that should be optimized through a learning process. The aforementioned article has first trained a CNN and several traditional classifiers such as SVM and logistic regression to independently predict the type (benign or malignant) of lung cancer. Predictions are consequently utilized to train a second-stage classifier to generate the final outcome. Any classifier such as SVM and NN can be used as the second-stage classifier.

2. Feature-level Fusion: Second widely used approach to combine deep and hand-crafted Radiomics is to first concatenate the feature vectors and then feed them to a classifier, referred to as feature-level fusion [66]. Reference [50] has shown that this combination lead to the highest performance in lung cancer survival prediction using Random Forest and Naive Bayes classifiers. The efficiency of this approach is also verified in [59] for lung tumor detection. Although mixing deep and hand-crafted Radiomics has several advantages such as ensuring the heterogeneity of the features, it may cause over-fitting as the number of training data is relatively less than the number of features. Therefore, Reference [68] has examined this large set of features in terms of stability, informativeness, and redundancy leading to a dramatic dimension reduction and increase in the accuracy of brain cancer survival prediction. To further reduce the number of features, a Cox regression is adopted that can determine the impact of features on survival, and as a result, those with small weights can be removed as effectless.

Reference [73] has leveraged the idea of concatenating deep and hand-crafted features for lung tumor attributes (such as spiculation, sphericity and malignancy) scoring through a multi-task learning framework. For extracting deep features, 9 CNNs corresponding to each of the 9 task at hand, and a 3 layer DAE are trained, where each CNN generates 192 Radiomics features extracted from the last fully connected layer before the SoftMax layer, and DAE results in 100 features. Deep features are further combined with hand-crafted features consisting of Haar and Histogram of oriented gradients (HoG) features, and the resulting vector is used as input to a multi-task linear regression model, which can consider the inter-task relations, in order to calculate the score of each of the 9 lung cancer attributes.

2.4 Conclusion

During the past decades, medical imaging made significant advancements leading to the emergence of automatic techniques to extract information that are hidden to human eye. Nowadays, the extraction of quantitative or semi-quantitative features from medical images, referred to as radiomics, can provide assistance in clinical care especially for disease diagnosis/prognosis. There are several approaches to radiomics including extracting hand-crafted features, and deep features. Furthermore, these two groups can be combined to take advantage of their inherited benefits and capabilities. We have presented an integrated sketch on radiomics by introducing practical application examples, and basic processing modules of the radiomics.

Chapter 3

Deep Learning-based Radiomics for Tumor Classification

This chapter considers the problem of tumor classification using deep learning-based radiomics by focusing on first brain tumor in Section 3.1 and then lung tumor in Section 3.2. As stated previously, tumor classification refers to determining the type of the tumor, where the focus is on three main classes: (i) Benign; (ii) Primary malignant, and; (iii) Metastatic malignant. While benign tumors, typically, do not spread to other organs but may need surgical resection as they may grow in size. Primary malignant tumors may not become metastatic for years, there is, however, possibility of transforming to aggressive malignant tumors, therefore, must be closely monitored/treated. The third category, i.e., the malignant tumors are life threatening and may spread to distant organs, requiring more complicated treatments. Consequently, tumor type classification using deep learning-based radiomics is of paramount importance.

3.1 Brain Tumor Classification

Brain tumor classification refers to determining the type of the tumor, having crucial impact on treatment design and selection. We begin this section by introducing an initial capsule network design, which is then improved by including the rough tumor boundaries. We, then, propose a boosting framework to facilitate the problem of architecture design. Finally, we model the prediction uncertainty through a Bayesian design of capsule network.

3.1.1 Brain Tumor Type Classification via Capsule Networks

According to cancer statistics, brain tumor is the leading cause of cancer death and it is among the most common cancers in children and adults [74]. Medical images are widely used for early detection of this cancer which leads to a more effective treatment. Among all medical imaging technologies, Magnetic Resonance Imaging (MRI) is more popular for brain tumor detection due to its harmless nature. Brain tumors have different types and determining these types for each patent is a crucial task, since it helps physicians to have a more precise treatment plan and predict the patient's response to the treatment [75]. In this work, we consider three types of brain tumors: Meningioma, Pituitary and Glioma. Tumor type classification by human inspection is a timely and error prone task, because it highly depends on the experience of the radiologist [76]. Due to this reason, nowadays, designing automated systems for brain tumor classification is of significance.

Most of the previous works on brain tumor classification consists of segmenting the tumor region from the MR images and then extracting different types of features to classify the tumors. Havaei *et al.* have provided one of the most recent works for brain tumor segmentation. This work has proposed a two path Convolutional Neural Network (CNN), which not only takes the pixel properties into account, but also considers the probabilities of neighbouring pixels [77]. After segmenting the tumor region, different types of engineered features can be extracted. Usman *et al.* have used intensity, intensity differences, neighbourhood and wavelet texture. These features are then used for the classification part using random forest classifier [78]. In [76], the effect of tumor region augmentation for three feature extraction methods is studied. These methods include intensity histogram, grey-level co-occurrence matrix (GLCM), and bag-of-words (BoW). Results of this paper show that the proposed criteria can enhance the accuracy of brain tumor classification. Abbadi *et al.* have also adopted GLCM and grey-level run length matrices(GLRLM) to extract 18 features for tumor classification using probabilistic neural network (PNN) [79].

All the aforementioned studies on tumor classification have a considerable drawback. They need a prior knowledge of kind of features to extract, which reduces their generalization capability. One of the most important advantages in machine

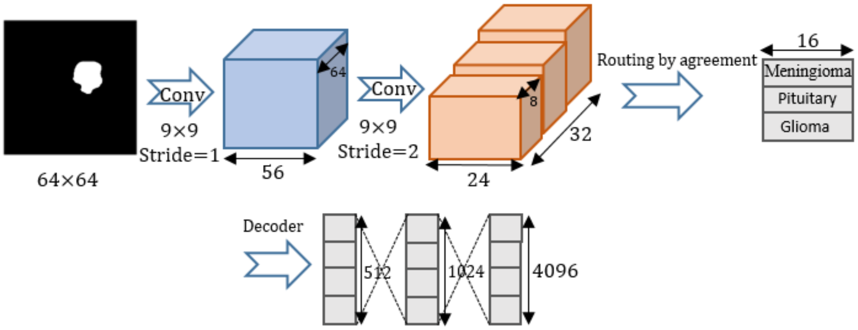


Figure 3.1: Proposed Capsule network architecture for brain tumor classification.

learning and specially vision tasks was the use of CNN [80]. These networks have a large learning capacity and can infer the image nature on their own without any prior knowledge, which makes them a suitable method for image classification. The use of CNNs for brain tumor type classification is explored in [81]. In this paper, neural networks and CNNs have been used with different kinds of preprocessings including data augmentation, and their results show that CNNs without any pre-processing outperform other methods on axial brain MR images. Although CNNs have successfully overcome many approaches in image processing, they still have some drawbacks. For instance, they are not robust to affine transformation and they do not take the spatial relationships into account. Therefore, they should be provided with data consisting of all kind of rotations and transformation to improve their generalization, and they perform poorly confronting small data sets, which is the case for most of the medical image databases, including brain MRI. Capsule networks can overcome these drawbacks [2]. In this sense, we begin solving the problem of brain tumor classification by developing a CapsNet-based architecture. The summary of the layers of our proposed model, shown in Fig. 3.1, is as follows:

- Inputs to the model are MRI images which are down-sampled to 64×64 from 512×512, in order to reduce the number of parameters in the model and decrease the training time. We did not observe any performance degrade due to this sown-sampling. Later, in Section 3.1.2, we show how a boosting approach can relieve the need for exhaustive parameter space exploration.

- Second layer is a convolutional layer with $64 \times 9 \times 9$ filters and stride of 1 which leads to 64 feature maps of size 56×56.

- Next layer is a capsule layer which is the result of $256 \times 9 \times 9$ convolutions with strides of 2. This layer consists of 32 capsules with dimension of 8 and the size of feature maps are 24×24.

- final capsule layer includes 3 capsules, one for each type of tumor. The dimension of these capsules is 16.

- The decoder part is composed of three fully connected layers having 512, 1024 and 4096 neurons respectively. The number of neurons in the last fully connected layer is the same as the number of pixels in the input image, as the goal is to minimize the sum of squared differences between input images and reconstructed ones.

For the goal of tumor type classification, two types of images can be used as inputs. We can use either the whole brain tissue or we can first segment the tumor regions and use these regions as the inputs to the classification model. As stated in the capsule network original paper [51], capsules tend to model everything in the input image, thus they do not perform as good for images with miscellaneous backgrounds. Due to this fact, we expect our capsule network to have a better result when fed with segmented tumors instead of the whole brain images.

To test our proposed approach, we have used the data set presented from [43, 76]. This data set contains $3,064$ MRI images of 233 patients diagnosed with one of the brain tumor types. The most important property of this data set is that it includes both the brain images and the segmented tumors, which enables us to perform experiments on two types of inputs. The first part of our experiments is allocated to testing different kinds of capsule network architectures. We have changed different parts of the original framework and calculated the prediction accuracy. Table 3.1 shows the obtained results. According to these results, reducing the number of feature maps from 256 (as in the original architecture) to 64 leads to the highest accuracy.

After selecting the best architecture of the capsule network, we have compared its classification accuracy with a CNN. The CNN we have used is adopted from [81], which has investigated the problem of brain tumor classification on the same data set we have used. The layers of this CNN are as follows:

- Convolutional layer with $64 \times 5 \times 5$ filters and strides of 1

Table 3.1: Brain tumor classification accuracy based on different capsule network architectures.

Capsule network architecture	Prediction accuracy
Original architecture	82.30%
Two convolutional layers with 64 feature maps each	81.97%
One convolutional layer with 64 feature maps	**86.56**%
One convolutional layer with 64 feature maps and 16 primary capsules	83.61%
One convolutional layer with 64 feature maps and 32 primary capsules of dimension 4	82.30%
Three fully connected layers with 1024, 2048 and 4096 neurons	83.93%

- 2×2 Max-Pooling
- Convolutional layer with $64 \times 5 \times 5$ filters and strides of 1
- 2×2 Max-Pooling
- Fully connected layer of 800 neurons
- Fully connected layer of 800 neurons
- Fully connected layer of 3 neurons

We compared capsule network with the CNN for both brain images and segmented tumors. Obtained results is shown in Fig. 3.2. Based on this figure, Capsnet outperforms CNN for both types of inputs. As it is stated in the original Capsnet paper, capsules tend to account for everything in the input image even in the background, and considering the fact that brain MRI images are taken from different angles such as Sagital and Coronial, backgrounds have lots of variations. Therefore, Capsnet cannot handle brain images as good as segmented tumor images, and this may be one of the reasons Capsnet results in lower accuracy for brain images than for the segmented tumor ones.

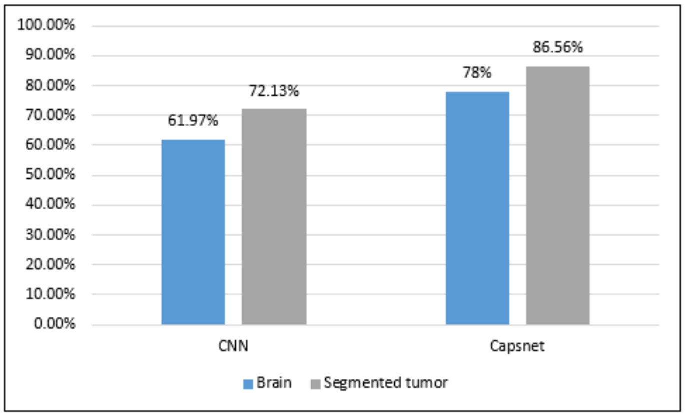


Figure 3.2: Capsnet and CNN accuracy for brain and segmented tumors images.

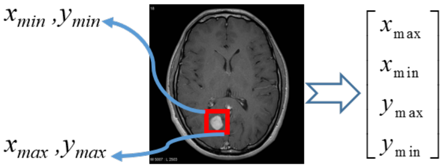

x_{min}, y_{min}
x_{max}, y_{max}
$\begin{bmatrix} x_{\max} \\ x_{\min} \\ y_{\max} \\ y_{\min} \end{bmatrix}$

Figure 3.3: Defining the tumor boundary box.

Capsule Networks with Coarse Tumor Boundaries

Although CapsNet has a better performance on the segmented tumor, needing the segmented tumor has two major problems: (i) First, segmenting the tumor is a time-consuming task and can only be provided by experts, and; (ii) Second, the tumor surrounding tissue contains valuable information, which is not accessible, when the network is fed with only the segmented region.

We address the aforementioned issues by *giving CapsNet the access to the tumor surrounding tissues, without distracting it from the main target, and requiring the tumor detailed annotation* [3]. More specifically, to help the CapsNet to focus on the main region while, at the same time, use the information from the surrounding tissues, we provide the network with the tumor coarse boundaries, leading to a modified CapsNet architecture, referred to as “BoxCaps”.

BoxCaps Architecture: The vector containing the tumor boundary, shown in the Fig. 3.3, is concatenated with the output of the capsule layer, and goes through a set of fully connected layers, in order to make the final decision, which is the type of the tumor. The detail of the proposed architecture is as follows:

- The inputs to the network are brain MRI images which are downsampled to 128×128 from 512×512.
- Second layer is a convolutional layer, with a total of 64 feature maps. The size of the filters is 9×9 with stride one.
- The third layer is a capsule layer resulted from 9×9 convolutions. This layer contains 32 capsules of dimension 8.
- The last capsule layer, which contains one capsule for each brain tumor type, determines the most probable class, along with its instantiation parameters. Outputs from this layer are masked based on the detected class, i.e., all capsules, but the winner, are set to 0.
- The tumor boundary box is concatenated with the obtained masked vector and goes through two fully connected layers, with 512 and 1024 neurons, respectively.
- The last layer is a Softmax layer that outputs the probability of each class being present.

Loss Function: The loss for the output of the capsule layers should be added to the Softmax layer loss, which we have defined as a cross entropy loss, with a smaller weight (indicated in Table 3.2), not to dominate the final loss. As such, we have defined the final loss as

$$\text{Loss} = \gamma \underbrace{\times \sum_{j=1}^{K} l_j}_{\text{Capsule Loss}} \underbrace{- \sum_{j=1}^{K} y_j \log\big(p(y_j)\big)}_{\text{Cross Entropy Loss}}, \tag{3.1}$$

where y_j is a binary variable indicating whether class j is present or not. Term $p(y_j)$ is the probability of this class being present, which is determined by the network, and; K is the number of output classes (types of the tumor). This loss is back propagated through the whole network, including both capsule and fully connected layers.

As shown in Table 3.3, the proposed BoxCaps architecture is compared with different alternative scenarios where the network is fed with either the brain or the segmented tumor image. In Table 3.3, we have also included the result for a modified CNN adapted based on the proposed architecture. The modified CNN takes as input

Table 3.2: Training hyper-parameters used for brain tumor classification via Adam optimizer.

Hyper-parameter	**Optimized Value**
Optimizer	Adam
Number of Epochs	50
Batch size	16
Routing iteration	3
Learning rate	0.01
Learning rate decay	0.9
γ (in Eq. (3.1))	0.1
λ (in Eq. (2.12))	0.5
m^+ (in Eq. (2.12))	0.9
m^- (in Eq. (2.12))	0.1

Table 3.3: Comparison between the proposed CapsNet BoxCaps and previous results. The bold number corresponds to the proposed approach, which outperforms its counterparts.

	Approach	**Accuracy**
1.	CapsNet given brain image as input [2].	78%
2.	CapsNet given segmented tumor as input [2].	86.56%
3.	Proposed BoxCaps Architecture.	**90.89%**
4.	CNN given brain image as input [81].	61.97%
5.	CNN given segmented tumor as input [81].	72.13%
6.	Modified CNN with brain image and tumor boundary box as inputs.	88.33%

both brain images and bounding boxes, where the box coordinates are concatenated with the last fully connected layer of the CNN. As it can be inferred from Table 3.3, the CapsNet architecture introduced in our study outperforms CNN in all situations, and achieves the best performance when it is fed with brain images, and coarse tumor boundaries.

3.1.2 BoostCaps: A Boosted Capsule Network for Brain Tumor Classification

In Section 3.1.1, we showed that capsule networks considerably outperform CNNs in brain tumor type classification. However, we observed that these networks, being sensitive to miscellaneous backgrounds, have a better performance when being fed with the segmented tumor rather than the whole brain image. Segmenting the

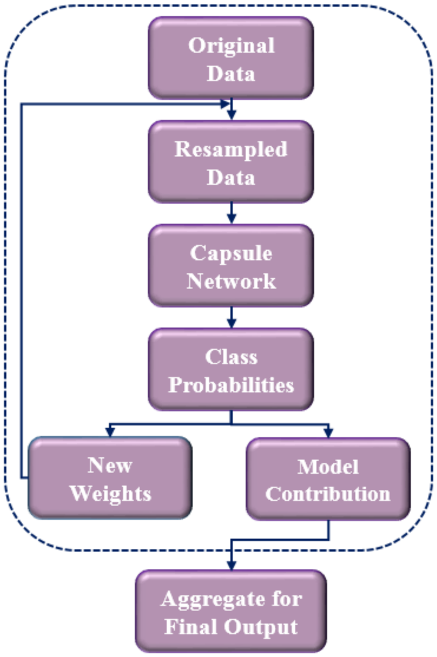


Figure 3.4: The proposed BoostCaps framework.

tumor, however, is time consuming and prone to inter-observer variability. More importantly, it completely loses the information about the location of the tumor and the surrounding tissue, which are of paramount importance in brain tumor classification. To tackle this problem, we developed the BoxCaps architecture that took not only the whole brain image, but also the coarse boundary box of the tumor as the input. This architecture, having the advantages of providing information about the whole brain tissue, and not requiring the exact segmented tumor, improved the accuracy of the brain tumor classification.

Capsule networks, similar to other deep learning networks, can have various architectures, depending on the number of layers, number of capsules, activation functions and many more. Exploring all these architectures, in order to find the most accurate one for the problem at hand, is significantly time-consuming and requires powerful computational resources. One possible solution to eliminate the need for searching in the space of all possible architectures is to take a boosting approach [82]. Boosting, which is a committee-based machine learning technique, starts with a weak learner (simple machine learning model) and trains this model, over and over again, by giving more weights to miss-classified samples, at each step. Accordingly, the final prediction is the weighted average of all predictions coming from the weak learners, where weights are determined based on how successful each learner has been in predicting the correct labels. In this study, we take the BoxCaps architecture proposed in Section 3.1.1, and adopt a boosting approach, namely the Stage-wise Additive Modelling using a Multi-class Exponential loss function. Real (SAMME.R) [83], to make the most out of the network, without the need to explore all possible architectures. Our proposed boosted capsule network, referred to as "BoostCaps" [4] is , to the best of our knowledge, the first capsule network that incorporates a boosting approach. Furthermore, boosting deep learning has been rarely applied to medical imaging problems, including the brain tumor classification.

Boosting [84] is a method for improving performance of machine learning techniques accomplished by combining several weak learners. These learners are trained over and over again, while at each round t more weight is assigned to certain samples, leading to the modified distribution, denoted by $\mathcal{D}_t$ hereafter, over the training set. Distribution $\mathcal{D}_t$ will then directly effect the cost function, or change the way training instances are sampled. Generally speaking, in boosting techniques, weights are distributed such that incorrectly classified instances, will be associated with higher weights, and higher chance of being picked, accordingly. Generally speaking, there are the following three different training strategies available in a boosting approach:

1. The t^{th} classifier is trained on instances resampled from the original data, with respect to distribution $\mathcal{D}_t$. This resampled dataset is used in all the epochs.

2. For training the t^{th} classifier, different resampled instances are used for each and every epoch.

3. The t^{th} classifier is trained by directly weighting the cost function, where miss-classified samples are associated with higher weights.

In deep learning applications, the first two criteria are more common.

BoostCaps Framework

The developed framework, referred to as a boosted capsule network (BoostCaps), shown in Fig. 3.4, is summarized in Algorithm 1. Notations used in this algorithm are as follows: w_i is the weight associated with the i^{th} sample, N denotes the total number of samples; M is the total number of capsule networks (weak learners); $T^m(x)$

```
Algorithm 1: BoostCaps
Result: The tumor type $C(x)$
Initialize the observation weights $w_i = 1/N$;
for $m = 1$ to $M$ do
    (a) Fit a capsule network $T^m(x)$ to the training data resampled based on
     the weights $w_i$;
    (b) $P_k^m(x) = \text{Prob}(c = k|x)$, $k = 1, ..., K$;
    (c) $h_k^m(x) = (K-1)\Big(\log P_k^m(x) - 1/K \sum_{k'} \log P_{k'}^m(x)\Big)$, $k = 1, ..., K$;
    (d) $w_i = w_i.exp\Big(-\frac{K-1}{K} y_i^T \log P^m(x_i)\Big)$, $i = 1, ..., n$;
    (e) Re-normalize $w_i$;
end
$C(x) = \underset{k}{argmax} \sum_{m=1}^{M} h_k^m(x)$;
```

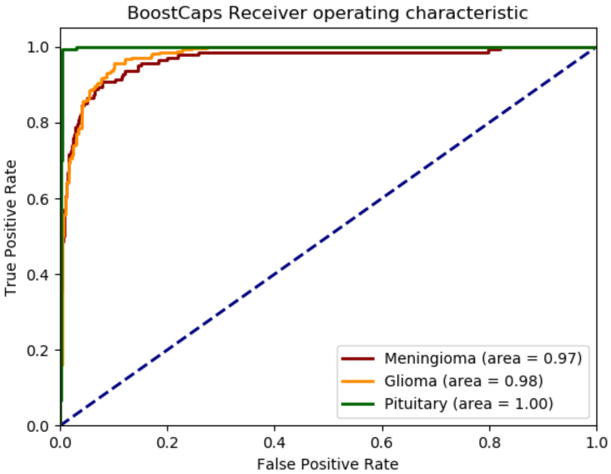


Figure 3.5: ROC curve for the three tumor types predicted by the proposed BoostCaps.

represents the m^{th} capsule network that takes a sample x as input; $P_k^m(x)$ is the probability of the sample x belonging to the class k, for $(1 \leq k \leq K)$, based on the m^{th} capsule network; $h_k^m(x)$ is the contribution of the m^{th} capsule network to the class k for the sample x; y_i is the true label for the i^{th} sample; Symbol T denotes the transpose operation, and; $C(x)$ is the final output of the proposed BoostCaps.

To test the proposed BoostCaps framework, shown in Fig. 3.4, we used the brain cancer dataset introduced in Section 3.1.1, where 20% of the data is set aside for testing purposes. We have trained our proposed BoostCaps framework, using 10 capsule networks as the weak learners. Although the number of the weak learners is

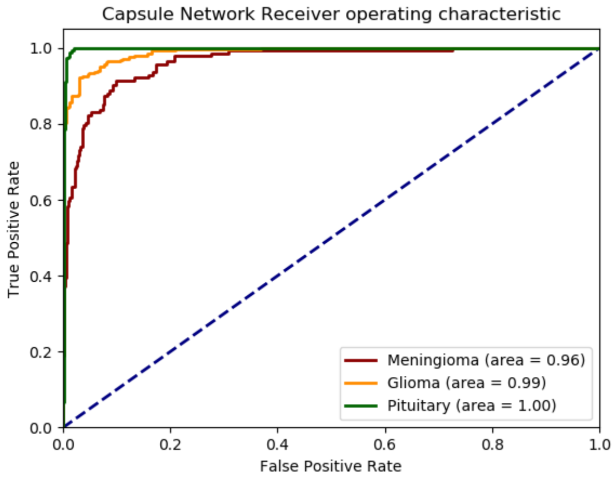


Figure 3.6: ROC curve for the three tumor types predicted by the capsule network.

usually much larger than 10, a deep learning model can bring training error to zero after a few steps [84]. Each capsule network is trained for 100 epochs, with a batch size of 16, and 3 routing iterations. We compared the proposed BoostCaps with a single BoxCaps model, in terms of accuracy, sensitivity, specificity, and area under the curve (AUC), where all the metrics, except accuracy, are calculated for the three classes separately. Obtained results are presented in Table 3.4. As it can be inferred from this table, the proposed BoostCaps can outperform the capsule network in terms of accuracy, sensitivity for the first and the third class, specificity for all three classes, and AUC for the first class. The receiver operating characteristic (ROC) curves is also presented in Figs. 3.5 and 3.6, for the BoostCaps and the capsule network model, respectively. Based on these figures, while BoostCaps results in relatively similar performance for the first two classes, there is a gap between them, in the case of using a single capsule network.

3.1.3 BayesCap: A Bayesian Approach to Brain Tumor Classification Using Capsule Networks

Similar to other standard deep learning networks, CapsNets do not capture model uncertainty [85], referring to how much the model is uncertain about its weights and thus the predictions. Measuring this uncertainty is critical as it provides a means to

Table 3.4: Results obtained from the proposed BoostCaps and the capsule network.

	BoostCaps	**Capsule Network**
Accuracy	**92.45%**	89.83%
Sensitivity for Meningioma	**75.35%**	64.79%
Sensitivity for Glioma	96.85%	**97.89%**
Sensitivity for Pituitary	**98.9%**	96.72%
Specificity for Meningioma	**97.64%**	97.43%
Specificity for Glioma	**88.61%**	82.77%
Specificity for Pituitary	**89.69%**	82.77%
AUC for Meningioma	**0.97**	0.96
AUC for Glioma	0.98	**0.99**
AUC for Pituitary	0.99	0.99

return the uncertain predictions to experts, and develop a human-in-the-loop mechanism. In other words, uncertainty estimates have the promise of improving: (i) The time-efficacy, by keeping and processing certain inputs, and; (ii) The accuracy, referring the uncertain ones to human experts [85]. In a recent study by Kendall and Gal [86], several examples are provided illustrating the importance of modeling uncertainty, without which disastrous mistakes can happen.

Most of the deep learning models, developed for classification problems, are associated with a softmax output that assigns probabilities to possible classes. The softmax output (similar to the squashing function in CapsNets), however, is often mistakenly interpreted as a measure of the model's confidence about its prediction [85]. Nevertheless, it has been shown that a model can have a high softmax output for its uncertain predictions. Motivated by this, Bayesian theory provides a mathematically grounded solution to model the uncertainty associated with model weights. Bayesian deep learning has drawn significant attention [87–89] recently. Arming a deep learning network with Bayesian reasoning often reduces the accuracy and is associated with higher computational cost [85]. However, it paves the path for measuring the prediction uncertainty, keeping the human in the loop, and improving overall interpretability of the network, which are critical for medical applications.

Returning the uncertain inputs to the human experts is based on the hypothesis that uncertain predictions tend to be incorrect [90]. One possible approach to test this hypothesis is to define an uncertainty index, such as the sample variance of the predictions made from different forward passes through the network [91], and investigate the association between this index and the incorrect predictions. If removing the uncertain predictions can enhance the performance, uncertain inputs are in fact the incorrect ones, and it seems reasonable to refer them to the experts for subsequent clinical revisions [90]. Capitalizing on this intuition, Nair *et al.* [90] have proposed several uncertainty measures in deep learning that can be correlated to incorrect predictions. Their results indicate that the uncertainty measures are indeed useful for human in the loop construction. Ozdemir *et al.* [92] have, additionally, studied the role of uncertainties in improving the network performance by propagating the uncertainty through the pipeline leading to enhanced performance, in terms of accuracy and confidence of the lung tumor detection.

Here, we propose a Bayesian CapsNet [5], referred to as the BayesCap, for the task of brain tumor type classification, which differs from the previous studies from different aspects. Derived from Variational Bayes, Reference [93] proposed a new Capsule routing algorithm that does not consider a probability distribution over all the model weights to capture model uncertainty. Reference [94], on the other hand, replaces the Bayesian framework with a Dropout procedure, through which, neurons are activated randomly during both training and testing. This means that rather than considering a trainable probability distribution over the model weights, a predefined Bernoulli function is utilized. The proposed BayesCap benefits from: (i) Capsule network architecture design being capable of handling small datasets, while inferring the spatial relations; (ii) Bayesian inference that defines priors over the network parameters, and hence managing the over-fitting of the model by avoiding the need to rely on point estimations alone; (iii) A means to asses the prediction uncertainty by leveraging a Monte Carlo approach, and; (iv) An index to measure the uncertainty and filtering out the uncertain predictions. The proposed BayesCap can be a part of a larger framework to accelerate the clinical decision-making, instead of completely replacing the human-centered procedures. Besides accuracy and time-related advantages, as deep learning is often treated as a black-box method, the proposed approach can improve the trust in such techniques.

As shown in Fig. 3.7, the following formalizes the BayesCap architecture, taking advantage of the ability of the capsule networks to handle small datasets and the Bayesian framework to manage the uncertainty.

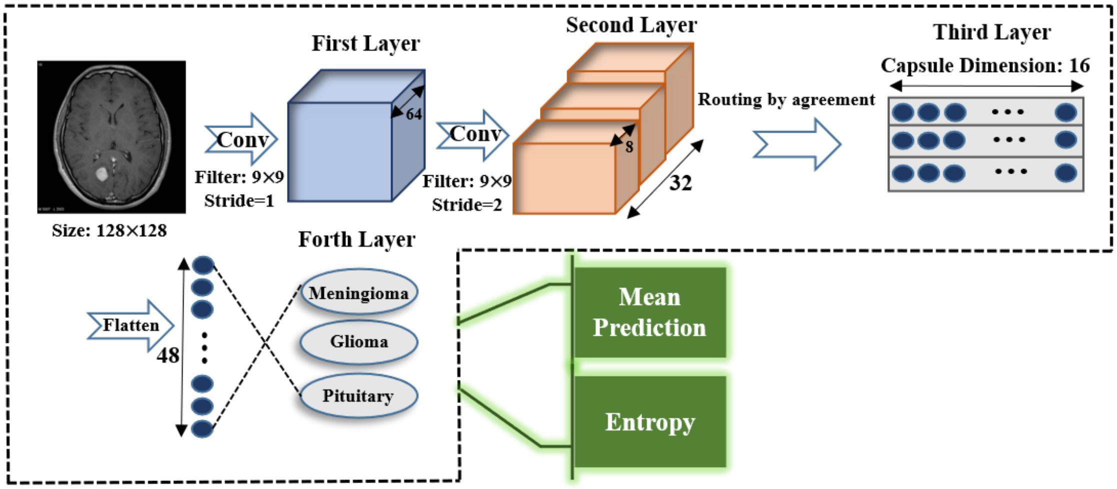


Figure 3.7: Proposed BayesCap architecture, capable of outputting the mean prediction, as well as the prediction entropy.

Bayesian Formulation of Capsule Networks

To develop the proposed BayesCap framework, without loss of generality, let's consider two capsule layers, i.e., a primary capsule layer consisting of N_{Pr} capsules, and; a parent capsule layer with N_{Pa} capsules. Each Capsule i is a group of neurons, conveying information about the object they are representing. This information includes both the object instantiation parameters $\boldsymbol{u}_i$, and its probability of being present as represented by the length of the capsule ($\|\boldsymbol{u}_i\|$). Each Capsule i, for ($1 \leq i \leq N_{Pr}$), in the primary capsule layer predicts the output of a parent Capsule j, for ($1 \leq j \leq N_{Pa}$), using a predication weight matrix $\boldsymbol{W}_{ij}$ (trained through back propagation), as $\hat{\boldsymbol{u}}_{j|i} = \boldsymbol{W}_{ij}\boldsymbol{u}_i$, where $\hat{\boldsymbol{u}}_{j|i}$ denotes the prediction of Capsule i for Capsule j. Next and before presenting the Bayesian agreement process of the proposed BayesCap, we formulate Bayesian modelling of the weight matrices ($\boldsymbol{W}_{ij}$) to form Capsule predications ($\hat{\boldsymbol{u}}_{j|i}$).

Bayesian Modelling of Predication Weight Matrices: All the network weights, including the prediction weights in the Bayesian agreement process ($\boldsymbol{W}_{ij}$), are defined as distributions to evolve through the back-propagation. To learn the predication weight matrices, we formulate the proposed BayesCap as a probabilistic model by defining prior and posterior distributions, denoted by $p(\mathcal{W})$ and $p(\mathcal{W}|\boldsymbol{\mathcal{D}})$, over the model weights $\mathcal{W} = [\boldsymbol{W}_{ij}]$, where $\boldsymbol{\mathcal{D}} = \{\boldsymbol{x}^{(n)}, y^{(n)}\}_{n=1}^{\mathcal{N}}$ is the training dataset consisting of $\mathcal{N}$ training instances, i.e., $\boldsymbol{x}^{(n)}$, for ($1 \leq n \leq \mathcal{N}$) and its associated label $y^{(n)}$.

By marginalizing over the predication weight matrices, the posterior predictive distribution $p(y|\boldsymbol{x}, \boldsymbol{\mathcal{D}})$ for each $\{\boldsymbol{x}^{(n)}, y^{(n)}\}$ can be defined as follows

$$p(y^{(n)}|\boldsymbol{x}^{(n)}, \boldsymbol{\mathcal{D}}) = \int p(y^{(n)}|\boldsymbol{x}^{(n)}, \mathcal{W})p(\mathcal{W}|\boldsymbol{\mathcal{D}})d\mathcal{W}. \tag{3.2}$$

Since computation of $p(\mathcal{W}|\boldsymbol{\mathcal{D}})$ is analytically intractable, it is approximated by a variational distribution $q(\mathcal{W}|\boldsymbol{\theta})$, having a known functional form parameterized by $\boldsymbol{\theta}$. Parameters of the variational distribution ($\boldsymbol{\theta}$) are estimated based on its statistical distance measure from the true distribution $p(\mathcal{W}|\boldsymbol{\mathcal{D}})$. To achieve this goal, different distance measures, such as Wasserstein [95] and Bhattacharyya [96] can be used. Here, for tractability of the derivations, we consider minimization of the Kullback-Leibler (KL) divergence between the two underlying distributions, i.e., $p(\mathcal{W}|\boldsymbol{\mathcal{D}})$ and $q(\mathcal{W}|\boldsymbol{\theta})$ to compute parameter $\boldsymbol{\theta}$. The cost function, $\boldsymbol{F}(\boldsymbol{\mathcal{D}}, \boldsymbol{\theta})$, therefore, is defined as follows

$$\begin{aligned} \boldsymbol{F}(\boldsymbol{\mathcal{D}}, \boldsymbol{\theta}) &\triangleq KL\big(q(\mathcal{W}|\boldsymbol{\theta})||p(\mathcal{W}|\boldsymbol{\mathcal{D}})\big) \\ &= \;\; KL\big(q(\mathcal{W}|\boldsymbol{\theta})||p(\mathcal{W})\big) - \mathbb{E}_{q(\mathcal{W}|\boldsymbol{\theta})} \log p(\boldsymbol{\mathcal{D}}|\mathcal{W}), \end{aligned} \tag{3.3}$$

where the equality is established following some simplifications, not included here to save on space. The cost function (Eq. (3.3)) can be re-written as follows

$$\begin{aligned} \boldsymbol{F}(\boldsymbol{\mathcal{D}}, \boldsymbol{\theta}) = \mathbb{E}_{q(\mathcal{W}|\boldsymbol{\theta})} \log q(\mathcal{W}|\boldsymbol{\theta}) - \mathbb{E}_{q(\mathcal{W}|\boldsymbol{\theta})} \log p(\mathcal{W}) \\ - \mathbb{E}_{q(\mathcal{W}|\boldsymbol{\theta})} \log p(\boldsymbol{\mathcal{D}}|\mathcal{W}). \end{aligned} \tag{3.4}$$

Consequently, through a Monte-Carlo simulation and sampling from $q(\boldsymbol{w}|\boldsymbol{\theta})$, the cost function can be approximated as

$$\begin{aligned} \boldsymbol{F}(\boldsymbol{\mathcal{D}}, \boldsymbol{\theta}) \simeq \frac{1}{N_M} \sum_{m=1}^{N_M} \Big[\log q(\mathcal{W}^{(m)}|\boldsymbol{\theta}) - \log p(\mathcal{W}^{(m)}) \\ - \log p(\boldsymbol{\mathcal{D}}|\mathcal{W}^{(m)}) \Big], \end{aligned} \tag{3.5}$$

where N_M is the number of Monte Carlo iterations. In practice, through the training phase, the first two terms of the cost function $\boldsymbol{F}(\boldsymbol{\mathcal{D}}, \boldsymbol{\theta})$ in Eq. (3.5) are analytically calculated, whereas the log-likelihood term ($\mathbb{E}_{q(\mathcal{W}|\boldsymbol{\theta})} \log p(\boldsymbol{\mathcal{D}}|\mathcal{W})$) is approximated by drawing a sample from the variational distribution. Commonly, the variational distribution ($q(\mathcal{W}|\boldsymbol{\theta})$) is modelled as a Gaussian with parameters $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\sigma})$. Then, in the backward-pass, $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ are updated. At the test time, different predictions are obtained by sampling from the variational distribution.

Bayesian Agreement Process: Similar to our Bayesian modelling of the prediction weight matrices ($\mathcal{W} = [\boldsymbol{W}_{ij}]$) to form capsule prediction vectors ($[\hat{\boldsymbol{u}}_{ij}]$), we formulate a Bayesian agreement step. During this step, the coupling coefficient (c_{ij}) is calculated based on which the output of the parent Capsule j, denoted by $\boldsymbol{s}_j$, is formed.

The coupling coefficient c_{ij} determines the score parent Capsule j assigns to $\hat{\boldsymbol{u}}_{j|i}$. Consider an auxiliary binary variable z_{ij} taking the value of one, when Capsules i and j are coupled, and zero otherwise. The posterior over z_{ij} is given by

$$p(z_{ij} = 1|\boldsymbol{s}_j, \hat{\boldsymbol{u}}_{j|i}) \approx p(z_{ij} = 1|\hat{\boldsymbol{u}}_{j|i}) \times p(\boldsymbol{s}_j|\hat{\boldsymbol{u}}_{j|i}, z_{ij}), \tag{3.6}$$

which is the product of the prior over z_{ij} and the output's likelihood. To form the coupling coefficients $[c_{ij}]$, the prior probability of z_{ij} given the prediction $\hat{\boldsymbol{u}}_{j|i}$ is denoted by term $\exp(b_{ij})$ and term $\exp(a_{ij})$ is used to represent the likelihood of $\boldsymbol{s}_j$ given the prediction and z_{ij}. Consequently, the posterior in Eq. (3.6) reduces to

$$p(z_{ij} = 1|\boldsymbol{s}_j, \hat{\boldsymbol{u}}_{j|i}) \approx \exp(b_{ij}) \times \exp(a_{ij}). \tag{3.7}$$

Term a_{ij} is estimated by measuring the similarity between each prediction $\hat{\boldsymbol{u}}_{j|i}$ of Capsule j and its actual output $\boldsymbol{s}_j$. The aforementioned similarity, also referred to as the agreement coefficient, is calculated taking the inner product of the two underlying capsule vectors as $a_{ij} = \boldsymbol{s}_j.\hat{\boldsymbol{u}}_{j|i}$. The coupling coefficient c_{ij} is, accordingly, computed as

$$c_{ij} = \frac{p(z_{ij} = 1|\boldsymbol{s}_j, \hat{\boldsymbol{u}}_{j|i})}{\sum_l p(z_{il} = 1|\boldsymbol{s}_l, \hat{\boldsymbol{u}}_{l|i}))}, \tag{3.8}$$

where l is the possible capsules in the parent layer. Output of Capsule j is, finally, calculated by summing over all the predictions, taking the coupling coefficients into account, i.e.,

$$\boldsymbol{s}_j = \sum_i c_{ij}\hat{\boldsymbol{u}}_{j|i}. \tag{3.9}$$

Through the Bayesian agreement process, prior ($p(z_{ij} = 1|\hat{\boldsymbol{u}}_{j|i})$) and posterior ($p(z_{ij} =

$1|\boldsymbol{s}_j, \hat{\boldsymbol{u}}_{j|i}))$ over z_{ij} repeatedly replace each other, for a pre-defined number of iterations.

Architecture of the BayesCap

Structure of the BayesCap is as follows:

- The first layer is a convolutional one with 9×9 filters and stride of 1, outputting 64 feature maps.
- The second layer, referred to as the primary capsule layer, is formed by 9×9 convolutions with stride of 2. The resulting feature maps are reshaped to generate 32 capsules of dimension 8.
- The third layer, referred to as the main capsule layer, is formed through the routing by agreement process, leading to 3 capsules of dimension 16.
- The last layer is a fully connected layer that produces the final class probabilities, through a softmax activation.

One important aspect of the Bayesian framework is that the model should notify the human experts, in case of uncertain predictions. An uncertainty index is, therefore, required to determine such predictions. In this study, we adopt "Entropy" as an uncertainty measure, defined as

$$\hat{y}_k = \frac{1}{T}\sum_{t=1}^{T} \hat{y}_{k,t}, \qquad \text{and} \qquad \mathcal{H} = -\frac{1}{K}\sum_{k=1}^{K} \hat{y}_k \log(\hat{y}_k), \tag{3.10}$$

where T is the total number of samples drawn for the same input, and $\mathcal{H}$ is the entropy of the model predictions over the underlying input. K is the number of output classes, $\hat{y}_{k,t}$ is the estimated probability of class k in the t^{th} drawn of the model weights, and $\hat{y}_k$ is the final estimation for the class k after T iterations. In other words, the model's uncertainty about its prediction can be measured using this index to provide uncertain predictions to the human expert for further analysis. Eventually, at the test time, our proposed BayesCap can output both the mean prediction, by averaging over all the predictions, and entropy as its uncertainty. For further illustration, Algorithm 2 summarizes how the proposed model works at test time, using T Monte Carlo steps.

```
Algorithm 2: Bayesian CapsNet at test time
  Result: Tumor type and prediction uncertainty
  Feed instance $\boldsymbol{x}^{(n)}$ to the network;
  for $t = 1$ to $T$ do
      (a) Draw a sample from $q(\boldsymbol{w}|\boldsymbol{\theta})$;
      (b) Compute the tumor type probabilities, based on the weights obtained
        in previous step;
      (c) Store the probabilities obtained in the previous step;
  end
  Compute the tumor type by calculating the mean of the $T$ previously
    obtained predictions;
  Compute the prediction entropy, as a measure of uncertainty, using
    Eq. (3.10);
```

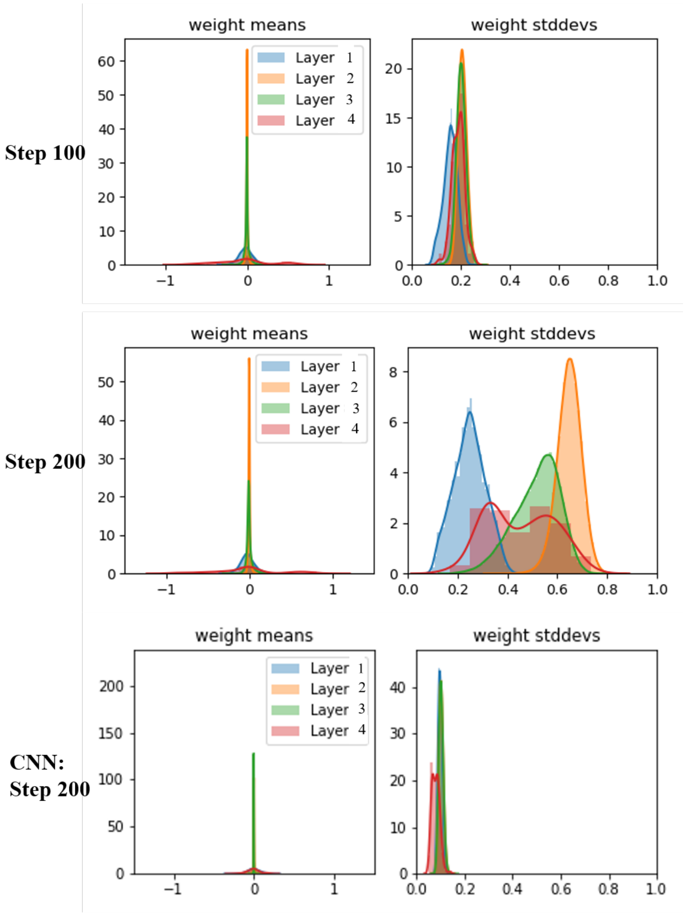
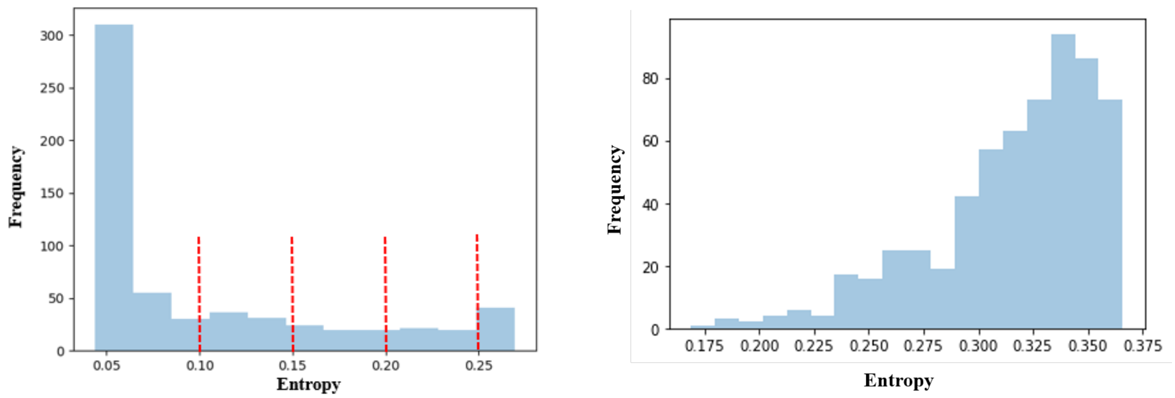


Figure 3.8: Evolution of BayesCap and CNN weights mean and standard deviation.

BayesCap Experiments

At the test time, the number of Monte-Carlo simulations [97] is set to 100 for calculation of the mean prediction. Fig. 3.8 shows the evolution of the weights' distributions obtained through the training phase. As it can be inferred from Fig. 3.8, while the distributions conform each other at the beginning, they tend to take distance, through the steps. The initial accuracy obtained after 500 epochs is 68.3%, which is higher than the accuracy of a non-Bayesian CNN trained on the same data (Section 3.1.1). This accuracy, however, is lower than 78% obtained from a non-Bayesian CapsNet using the whole brain image (Section 3.1.1), which is expected as the Bayesian version is trained with the goal of learning the posterior of model weights and capturing uncertainty rather than increasing accuracy.

As stated previously, one important advantage of modelling the uncertainty is to refer the uncertain predictions to the human expert for further follow ups. This process, however, is based on the hypothesis that the uncertain predictions are indeed the incorrect ones. To test this hypothesis, we calculated the uncertainty over the test set, and developed an uncertainty histogram, shown in Fig. 3.9a for one of the bootstraps. Based on this histogram, we filtered out predictions and their associated data sample, at several thresholds, shown in Fig. 3.9a. After filtering out the data, at each threshold, we calculated the prediction accuracy. We observed that starting from an accuracy of 68.3%, with no filtering, the performance gradually improves, at each threshold as follows: (i) By filtering out the predictions associated with an uncertainty over 0.25, where on average 65% percent of the uncertain predictions are incorrect, the accuracy improves to 71.3%; (ii) By removing the predictions with over 0.2 uncertainty, where on average 51% of the uncertain prediction are incorrect, the accuracy further improves to 72.6%, and; (iii) Finally, by filtering out the predictions associated with an uncertainty over 0.15 and 0.1, where on average 47% and 40% of the uncertain prediction are incorrect, the accuracy increases to 73.6%, and 73.9%, respectively. These observations, therefore, confirm the hypothesis that the uncertain predictions tend to be incorrect, and it makes sense to refer them to human experts, for further investigations. Finally, we trained a Bayesian CNN with the relatively same complexity (two convolutional layers and two fully connected layers) on the same dataset to compare the uncertainty in terms of entropy. As shown in Fig. 3.9b, the Bayesian CNN leads to a significantly higher prediction uncertainty, which means higher number of samples need to be referred to experts. Furthermore, last row of Fig. 3.8 presents the CNN weights' distributions after 200 steps, showing that they failed to adopt distinctive distribution parameters, compared to the BayesCap weights



(a) Uncertainty histogram for the test set in the BayesCap model.

(b) Uncertainty histogram obtained from a Bayesian CNN.

Figure 3.9: Uncertainty histograms.

provided in the same figure.

One important aspect of the Bayesian framework is the trade-off between exploiting/consulting the developed model and referring the samples to the experts. In other words, if too many samples need to be referred, one may choose to completely cast aside the model. The proposed BayesCap, with entropy of 0.25 used as the threshold (Fig. 3.9a), leads to referring less than 8% of the test instances. In the Bayesian CNN (Fig. 3.9b) with the same entropy of 0.25, however, most of the test instances are above the threshold, which means the relative failure of the CNN-based system. It is worth mentioning that through the development of the BayesCap framework, we have tried to capture the uncertainty in the model itself, referred to as the epistemic uncertainty [86,98]. Such uncertainty can be further reduced by collecting more data samples. Aleatoric uncertainty, on the other hand, is a different approach trying to calculate the inherited uncertainty of the data instances, caused by the observation noise. Including the aleatoric uncertainty in the proposed BayesCap is the focus of our upcoming research, through including the observation noise parameters in the loss function.

This completes developments and discussions on different CapsNets designs for the task of brain tumor type classification based on MRI images. Next, we shift the focus to lung nodule type classification based on CT scans.

3.2 Lung Nodule Classification

This section considers the problem of lung nodule classification, referring to determining malignant and benign tumors. First, a 3D multi-scale capsule network is proposed that can consider features extracted from neighbouring tissues. Then, a mixture of experts framework is designed, where each expert is specialized on particular types of the lung tumor.

3.2.1 3D-MCN: 3D Multi-Scale Capsule Network for Lung Nodule Malignancy Classification

Lung cancer is ranked first worldwide in terms of mortality and is among the top three cancer types in terms of incidence. Lung cancer together with breast cancer lead worldwide in terms of the number of new cases with approximately 2.1 million diagnoses estimated in 2018. Lung cancer is also responsible for the largest number of deaths (1.8 million deaths, 18.4% of the total), with a low 5-year survival rate (18%) [74]. The high mortality rate of the lung cancer is mainly due to the fact that lung cancer is diagnosed at advanced stages [74], in more than half of the cases. In recent years, significant technological advancements in medical imaging, especially Computed Tomography (CT) scans, have improved the detection rate of the lung tumor [99]. Analyzing and interpreting these images, however, is time consuming [24], and subject to inter-observer variability. Furthermore, intrinsic tumor heterogeneity that can significantly contribute to the cancer diagnosis, may not be visible to the human eye. [100].

To address the problem of lung nodule diagnosis, we propose a 3D multi-scale CapsNet [6], referred to as 3D-MCN, shown in Fig. 3.10, which takes 3D patches of the nodules at three different scales as inputs, and classify the nodule's malignancy. The motivation behind the proposed multi-scale technique is that the the nodule morphological characteristics are not the only indicators of its malignancy, incorporation of information obtained from the surrounding tissue and vessels play a critical role in determining the type of the nodule. In brief, the proposed approach benefits from: (i) 3D inputs, which give the model access to 3D features of the nodule; (ii) Multi-scale inputs, helping the network assess the local and global features; (iii) The CapsNet capability, when encountering with small datasets, and; (iv) Not requiring the nodule

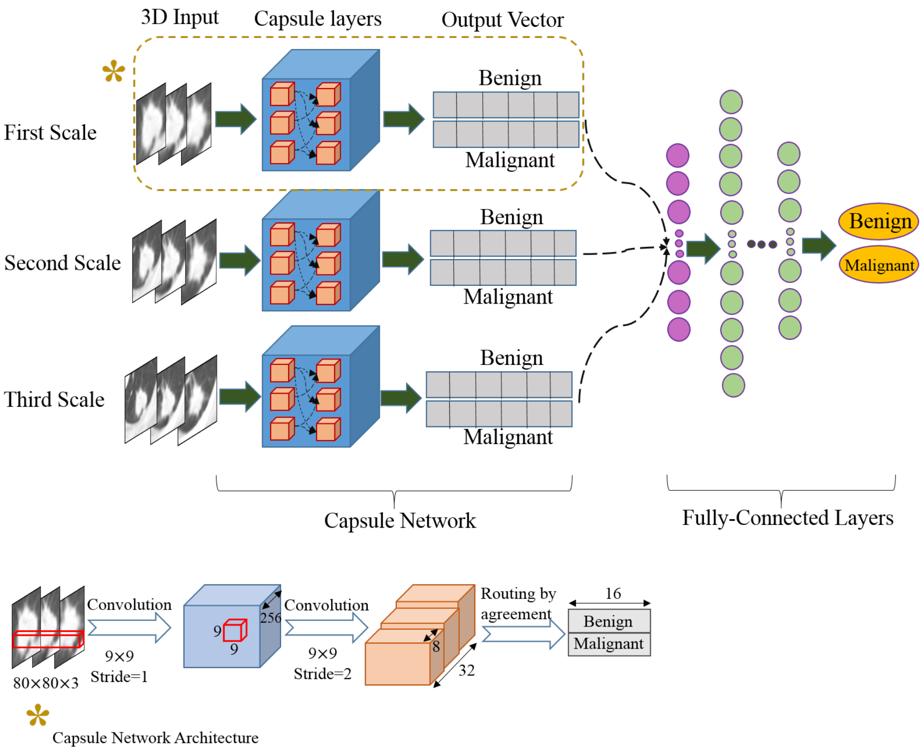


Figure 3.10: The proposed 3D-MCN framework. Three independent capsule networks take 3D nodule crops as inputs. Each CapsNet takes inputs at a different scale. The output vectors are masked and concatenated into a single vector. The resulted vector goes through a fusion module consisting of a set of fully connected layers to form probability associated with each class (benign or malignant).

detailed annotation and pre-defined features.

Dataset Description: We have conducted our experiments on the LIDC-IDRI [101–103], which is a collection of 1018 CT scans from 1010 patients. The nodules in this collection are identified and annotated through a two-phase process. In the first phase, 12 radiologists have independently reviewed the scans and marked the lesions as nodule≥ 3, nodule< 3, and non-nodule. Furthermore, the radiologists have annotated the ones identified as nodule≥ 3. In the second step, radiologists have had access to the results of the other radiologists to refine their own marks or leave them unchanged. After this phase, the radiologists have independently assessed several characteristics of the nodules≥ 3, including the likelihood of malignancy, shape, margin, and internal structure. Malignancy is rated from 1 to 5, where 1 indicates the lowest malignancy likelihood, and 5 denotes the highest. Fig. 3.11 shows illustrative examples of available

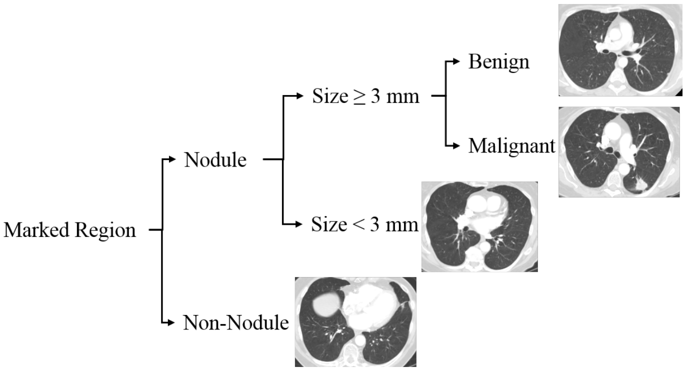


Figure 3.11: Examples of available marked regions in the LIDC-IDRI dataset. Each regions can be classified as nodule or non-nodule. Nodules can also be categorized based on their size. Nodules larger than 3 *mm* are further grouped based on their malignancy ratings.

marked regions in the dataset.

Nodule Patch Selection and Processing: In this study, we chose nodules$\geq$ 3 to classify them as benign (rating of 1 and 2) or malignant (rating of 4 and 5), based on the radiologists' provided ratings. We included all the marked nodules, in either the training or test set, even if the nodule was identified by only one radiologist, to have a model that is more robust to noisy inputs. The labels of the nodules that were identified by more than one radiologist, are the average over all the available ratings, rounded to the nearest integer. Consequently, nodules with an average malignancy of 3 (indeterminate malignancy) were discarded. For each nodule, we extracted three different 3D patches around the nodule center, where 3D patch refers to extracting one patch from the central slice, and two from the two immediate neighbors. As we later discuss in Section 4.2, selecting a higher number of channels leads to the following two important challenges: (i) First, it requires advanced memory resources, and; (ii) Second, it makes some tumors too small to be distinguished from surrounding tissues [143]. Furthermore, the 3-channel input has been previously investigated in several studies, leading to satisfying results. For instance, in Reference [144], 3-channel CT scans are used to predict short and long-term survival in lung cancer, using CNNs, and it has been shown that the 3-channel input outperforms the single channel one. The 3-channel input is also utilized in References [145] and [146], for classifying breast tumor and mediastinal lymph node metastasis of lung cancer, respectively, using CNNs. Each 3D patch was extracted at three different scales. The first scale completely fits the nodule boundary, based on the provided annotation. As nodules are associated with different sizes, all extracted patches were zero padded up to the fixed size of 80×80 (the biggest possible width and height based on the training data). The second scale was extracted by allowing a margin of 10 pixels at each side. The patches were zero padded to the fixed size of 100×100, and down-sampled to 80×80, to be consistent with the first scale, and reduce the complexity. Similarly, the third scale was extracted by allowing a margin of 20 pixels at each side. The patches were zero padded to the fixed size of 120×120, and down-sampled to 80×80. At the end, data was normalized between 0 and 1. The training set was shuffled and augmented by including random flipping. Finally we ended up having three sets of training (at three scales), three sets of test, one set of training label, and one set of test label. Each training set was fed to an independent CapsNet, along with its corresponding test set.

3D-MCN Architecture: Our multi-scale model is a fully connected neural network with 3 three hidden layers of sizes 1028, 512, and 256. The input to this network is the combination of all the output instantiation vectors from the three CapsNets. For each CapsNet, the output vector of the lower probability class is masked (set to zero). Each output class is of dimension 16. Having two output classes (benign and malignant) results in a vector of dimension 32 for each CapsNet, and having three CapsNets results in an input of size 96 to the multi-scale network. The output of this network is the probability of the nodule being benign or malignant, based on the information from all the three scales.

3D-MCN Experiments

Four measures of the area under the curve (AUC), accuracy, specificity, and sensitivity are calculated, based on the performance on the test set. The results are provided in Table 3.5, which shows that the proposed framework outperforms its counterparts in terms of the AUC, accuracy and sensitivity. Specificity is higher for the second scale model. However, we believe that sensitivity is of greater importance, as the consequences of miss-classification are worse for malignant cases. Furthermore, typically in clinical practice, suspected malignant cases will go over complementary examinations [104], which can identify whether the underlying case was a false positive. Fig. 3.12 illustrates the Receiver Operating Characteristic (ROC) curve for the single-scale models, as well as the multi-scale one, and the base-line, which is

Table 3.5: Performance of the proposed 3D-MCN approach along with performance of single scale-models, on the independent test set. The proposed approach outperforms others in terms of the AUC, accuracy and sensitivity.

Model	AUC	Accuracy	Specificity	Sensitivity
Proposed Model	**0.9641**	**93.12%**	90%	**94.94%**
First Scale	0.9633	91.65%	90%	92.21%
Second Scale	0.96	91.65%	**91.33%**	91.82%
Third Scale	0.96	91.40%	89.33%	92.60%
Base-line Model	0.9524	87.47%	86.66%	87.93%

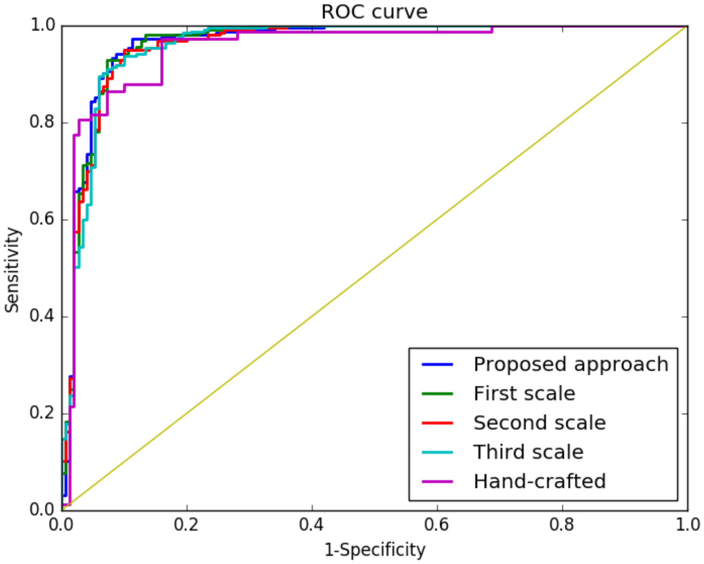


Figure 3.12: The ROC curve for the single-scale and multi-scale models, and the base-line model (i.e., fully connected layers trained on hand-crafted features only), showing that the proposed approach is capable of achieving higher AUC.

a fully-connected neural network (having the same architecture as the one in the multi-scale model) trained on four hand-crafted features, namely volume, diameter, center-of-mass x coordinate, and center-of-mass y coordinate, as the base-line model. These four features accompany the IDC-IDRI dataset, as means to ensure all research groups use the same size-selected nodules.

In clinical applications, where false positives and false negatives are not treated equally, the threshold resulting in the desired sensitivity and specificity can be selected based on the ROC curve. Another strategy to tune these measures is to assign different weights in the objective function. In this work, the proposed approach is

Table 3.6: Effect of changing the weights of the terms controlling the false positives and false negatives. As expected, it is observed that assigning more weight to the false positive loss, increases the sensitivity, while putting higher weight to the false negative loss, increases the specificity.

Weight Setting	**Area Under the Curve (AUC)**	**Accuracy**	**Specificity**	**Sensitivity**
Equal Weights	0.9641	93.12%	90%	94.94%
$\alpha = 2$ **and** $\beta = 1$	0.9641	93.12%	87.33%	96.49%
$\alpha = 1$ **and** $\beta = 2$	0.9638	92.87%	92%	93.38%

trained with the objective of minimizing the binary cross entropy loss [105, 106]. As such, we have modified the loss as follows to put different weights on the loss function terms (specificity and sensitivity)

$$loss = -(\underbrace{\alpha \;\; y \;\; \log(p)}_{\text{Controlling the false positives}} + \underbrace{\beta \;\; (1-y) \;\; \log(1-p)}_{\text{Controlling the false negatives}}), \tag{3.11}$$

where y is the target, and p is the predicted probability of Class 1. Terms α and β denote the weights given to the false positives and false negatives, respectively. We have trained the multi-scale model with three settings (equal wights, more weight assigned to the false positive, and more weight given to the false negative). Table 3.6 shows the obtained results.

We inspected the cases in the test set for which our proposed approach failed to predict the correct label. We realized that 28% of such failure cases are nodules that are marked by only one radiologist. In other words, there is no agreement on these cases being nodules between different radiologists. Although all other failure cases are nodules identified by at least two radiologists, there is a common pattern between most of them, i.e., the malignancy labels are not consistent, and moreover, there is at least one label 3, among the provided labels. In other words, although the average malignancy is not 3 to be discarded, there is a high chance that the underlying nodule is neither benign nor malignant.

As stated previously, the motivation behind our multi-scale approach is that the nodule morphological characteristics are not the only indicators of its malignancy. In fact, the surrounding tissue and vessels play an important role in determining in determining the benign or malignant nature of the nodule. [107]. To illustrate on the importance of having multi-scale inputs, we extracted the cases, where the output is different for different scales. Fig. 3.13 presents four nodules from three different

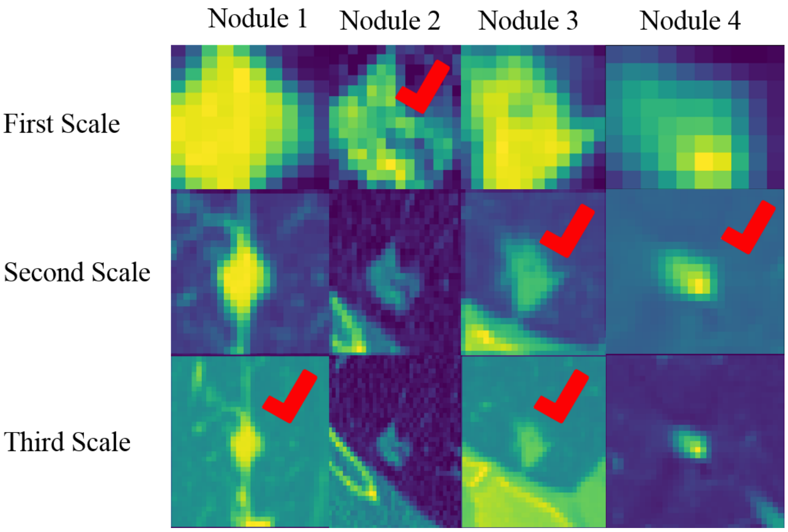


Figure 3.13: Cases, where not all the single-scale models provided correct predictions. The check sign indicates the successful scale. This figure illustrates the necessity of including all the three scales in the final model.

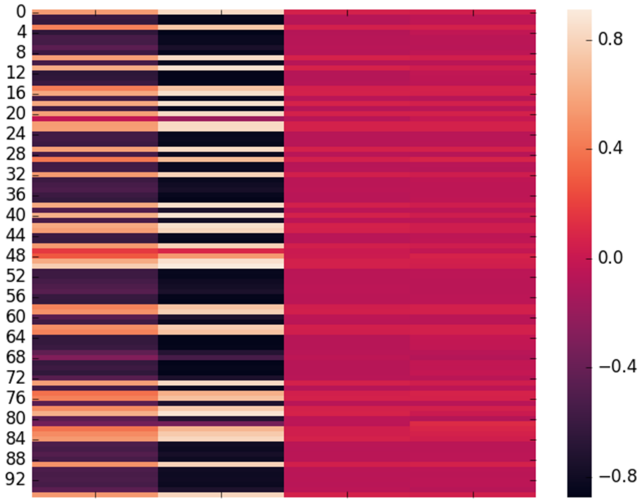
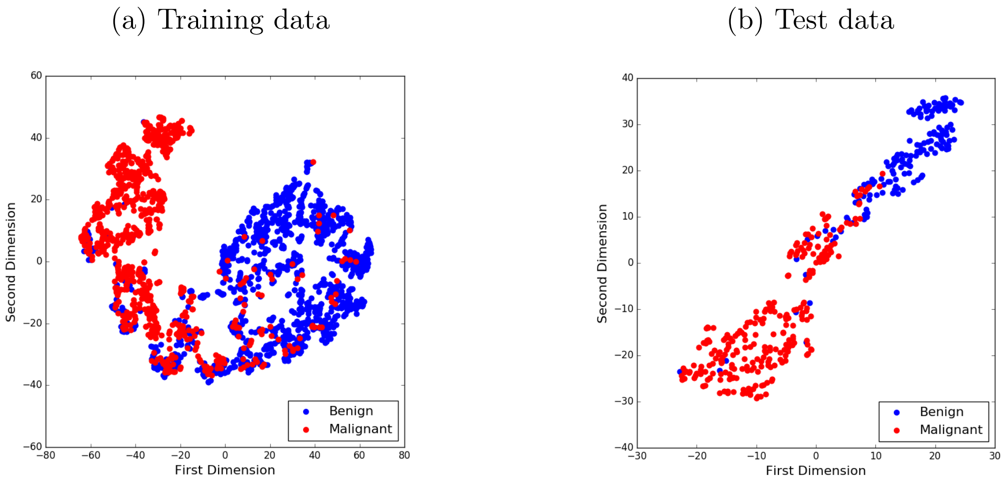


Figure 3.14: Correlation between the CapsNet features and the hand-crafted features. Most of the features are positively or negatively correlated with volume and diameter. However, the correlation with x and y centers is not considerable, as the model is fed with cropped nodule slices in different scales, and the location with respect to the whole image is not accessible to the model.

scales. The figure also indicates the scale which has been successful in classifying the nodule. In other words, having a correct prediction was not possible without including all the scales.



(a) Training data

(b) Test data

Figure 3.15: T-SNE plot of the CapsNet learned features in 2D, showing that the features are capable of distinguishing between the two classes.

In another experiment, we calculated the correlation between the CapsNet extracted features from all three scales with the four hand-crafted features of volume, diameter, x center-of-mass, and y center-of-mass, as shown in Fig. 3.14. Volume and diameter are important factors of the nodule malignancy, therefore, as expected, most of the learned features are highly (positively or negatively) correlated with these two features. The centers of mass are, however, calculated from the whole images, and as the model is only being fed with the cropped nodule slices via different scales, the learned features cannot represent these two characteristics.

Another important aspect that needs to be considered when extracting deep learning-based radiomics features is that weather they are capable of distinguishing the classes. We projected the high dimensional feature space of the CapsNet into a lower dimensional space, using a t-Distributed Stochastic Neighbor Embedding (t-SNE). The resulted feature space for both the training and test sets are shown in Fig. 3.15, according to which, features are distinctive, even in the simplified 2D space shown in Fig. 3.15.

There are two general paths to use the LIDC-IDRI dataset. The first one is to rely on the labels from the diagnosis data which is available for only 157 patients. The diagnosis labels are resulted from different examinations, including image review, biopsy, and surgical resection, at a nodule level. Such a pathway is explored in [108],

Table 3.7: A list of papers that have used LIDC-IDRI to predict the lung nodule malignancy based on the radiologists' provided ratings. Our proposed approach outperforms all the others except the third one. However, the third model incorporates hand-crafted features, needing exact annotations from the radiologists. Our proposed model, however, only requires nodule rough boundary box.

Method	**AUC**	**Accuracy**	**Specificity**	**Sensitivity**
Proposed model	0.9641	93.12%	90%	**94.94**%
CNN [100]	0.938	87.9%	87.9%	87.9%
CNN in combination with hand-crafted features [100]	**0.971**	**93.2**%	**98.5**%	87.9%
Deep residual network [110]	0.9459	89.90%	88.64%	91.07%
Deep belief network [25]	-	81.19%	-	-
CNN in combination with hand-crafted features [111]	-	86.79%	95.42%	60.26%
Multi-crop CNN [52]	0.93	87.14%	93%	77%

for lung nodule malignancy prediction, by training a CNN on the pathologically-proven diagnostic data. This approach has resulted in an accuracy of 77.52%, sensitivity of 79.06%, and specificity of 76.11%. Reference [109] is another example of using the diagnostic data, that has achieved an AUC of 0.981 and error rate of 5.41%.

The second path to use the LIDC-IDRI dataset (also followed in our work) is to adopt the ratings provided by experienced radiologists, at the time of reviewing the CT scans. In [100] a CNN based architecture is presented that can distinguish between benign and malignant nodules with an AUC of 0.938, accuracy of 87.9%, sensitivity of 87.9%, and specificity of 87.9%. The authors have further improved the performance to an accuracy of 93.2% by incorporating the hand-crafted features. A Random Forest (RF) classifier is trained on the combination of hand-crafted and deep learning-based features, to predict the nodule malignancy. Although the obtained accuracy is on a par with that of our proposed framework, it requires the nodules' fine annotations, from which our model is completely independent. Table 3.7 presents a list of papers that have used the same setting of the LIDC-IDRI as we did, along with their proposed method and obtained results.

One important challenge in comparing different studies on the LIDC-IDRI dataset is that different researchers have used different cohorts of training and testing. One solution to this challenge is to cross-validate the results, instead of using the fixed sets. This strategy, however, should be used with care, not to include nodules from the same patient in both sets. Another challenge of comparing the studies is the difference between reported performance measurements. Accuracy, which is the only metric provided in several studies, is not informative enough, as it gives no details on the portions of positive and negative samples, and a highly biased model can lead to a high accuracy. Our proposed Multi-scale model achieves high accuracy, as well as high sensitivity and specificity, showing that it is not biased towards positive or negative samples.

Another limitation of most of the previous models is that they require large number of samples and they rely on heavy data augmentation. However, the model proposed in this model is based on the CapsNet, which is capable of interpreting small datasets. This study shows that a good performance is achievable even in the lack of large datasets, which is typically the case in the medical domains.

3.2.2 MIXCAPS: A Capsule Network-based Mixture of Experts for Lung Nodule Malignancy Prediction

Capitalizing on the success of the CapsNets, we propose a new framework, referred to as the Mixture of Capsule networks (MIXCAPS), for the task of lung nodule malignancy prediction. The proposed MIXCAPS [7] framework is a "Mixture of Experts" type model [112–114], which has the potential to noticeably improve the classification accuracy by integrating/coupling several experts (individual CapsNets in the context of the proposed MIXCAPS). To be more precise, mixture of experts solves the classification problems by splitting the dataset into similar samples, and each expert specializes in classifying similar instances. To the best of our knowledge, the proposed MIXCAPS is the first CapsNet-based mixture of experts framework. The MIXCAPS model benefits from the following three important properties: (i) The embedded capsule network is capable of classifying the lung nodules without requiring availability of a large dataset; (ii) The mixture of experts approach enables each CapsNet within the MIXCAPS architecture to focus on a specific subset of the nodules, therefore, improving the overall classification performance of the model, and; (iii) As shown in our experiments, MIXCAPS is not restricted to the task of lung nodule malignancy prediction. In fact, it can be easily generalized to the prediction of other tumor types such as brain cancer.

Background: Mixture of experts (MoE) [112] refers to adopting several experts, each of which is specialized on a subset of the data, to collectively perform the final prediction task. As shown in Fig. 3.16, experts are separately fed with the input data and the final output is a weighted average of all the predictions coming from all the N active experts. The weight g_i assigned to Expert i can be either a pre-determined

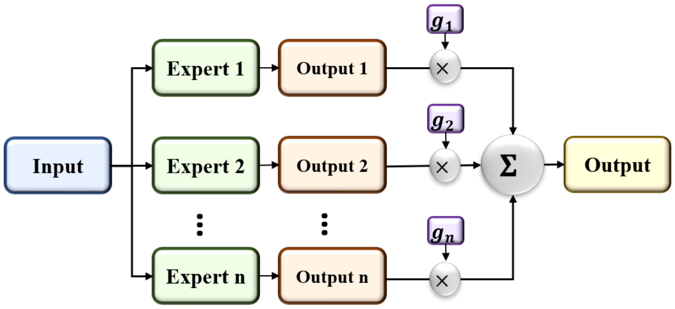


Figure 3.16: General framework of a mixture of experts approach.

value, or a trainable one. One simple example of the former case is averaging over all the experts' predictions [114]. However, more sophisticated approaches such as soft clustering of the input may also be adopted. In the latter case, weights may be trained at the same time with the experts. One other approach to use trainable gating weights is to concatenate the feature vectors obtained from the individual experts and feed the resulting vector to an external gating model to make the final decision.

The MoE concept has been widely used in medical imaging. The simple averaging scenario is investigated in References [115] and [116] for retinal vessel detection from fundus images and breast cancer detection from histology images, respectively. Trainable gating weights are studied in Reference [117], where hand-crafted and CNN-based features are combined to detect breast cancer from pathology images. The scenario where gating weights are trained at the same time with the experts is investigated in Reference [113] for breast cancer diagnosis. In particular, CNN experts are combined using weights coming from an external gating network. The gating network itself is a CNN, taking the same inputs as the experts, and outputting the probability of each expert being responsible for each particular input.

MIXCAPS Architecture

The proposed capsule network-based mixture of experts for lung nodule malignancy prediction, referred to as the MIXCAPS, is shown in Fig 3.17. The 3D nodule patches are the inputs to two capsule network experts, as well as the convolutional gating network. The two experts, as shown in Fig 3.17, consist of two convolutional layers, the last of which is reshaped to form a capsule layer. This capsule layer is followed by a routing by agreement and the final capsule layer. The outputs of the two experts, denoted by $\boldsymbol{o}_1$ and $\boldsymbol{o}_2$, represent the class (benign and malignant) probabilities. The

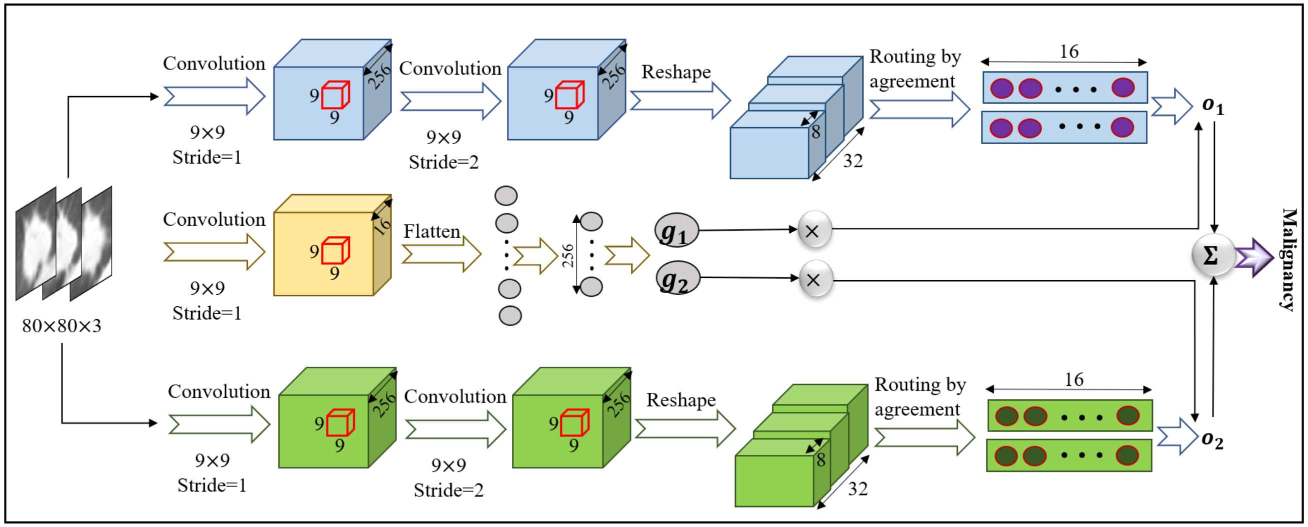


Figure 3.17: Proposed MIXCAPS.

gating network, consisting of a convolutional and two fully connected layers, determines the contribution of each expert, denoted by g_1 and g_2, for a specific input through a Softmax layer, as follows

$$g_1 = \frac{\exp(G_1)}{\exp(G_1) + \exp(G_2)}, \quad g_2 = \frac{\exp(G_2)}{\exp(G_1) + \exp(G_2)}, \tag{3.12}$$

where G_1 and G_2 are pre-activation outputs. The Softmax layer ensures that g_1 and g_2 sum to one. These contributions are multiplied by $\boldsymbol{o}_1$ and $\boldsymbol{o}_2$ to calculate the final prediction $\boldsymbol{o}$ as follows

$$\boldsymbol{o} = g_1\boldsymbol{o}_1 + g_2\boldsymbol{o}_2. \tag{3.13}$$

Output vector $\boldsymbol{o}$ encompasses the probability of benign and malignant classes, denoted by $o^{(0)}$ and $o^{(1)}$, respectively. In other words

$$\boldsymbol{o} = [o^{(0)}, o^{(1)}]^T. \tag{3.14}$$

where superscript T denotes transpose operator. Originally, margin loss is proposed for the training of the capsule networks. In this study, we adopt the same loss function with the difference that the loss l is calculated over the final output of the MIXCAPS instead of the individual capsule networks. In other words, The CapsNet loss function (margin loss) is modified to reflect the loss associated with the experts and gating models, as follows

$$l^{(0)} = T^{(0)} \max(0, m^+ - o^{(0)})^2 + \lambda(1 - T^{(0)}) \max(0, o^{(0)} - m^-)^2, \tag{3.15}$$

$$l^{(1)} = T^{(1)} \max(0, m^+ - o^{(1)})^2 + \lambda(1 - T^{(1)}) \max(0, o^{(1)} - m^-)^2, \tag{3.16}$$

$$l = l^{(0)} + l^{(1)}, \tag{3.17}$$

where $l^{(0)}$ and $l^{(1)}$ denote the losses associated with the benign and malignant classes, respectively. m^+, λ, and m^- are hyper-parameters. Terms $T^{(0)}$ and $T^{(1)}$ are the ground-truth labels for benign and malignant classes, respectively. According to Reference [112] comparing the desired output with the blend of outputs from the experts, leads to a strong coupling between experts and solutions in which many experts are used for one case. However, in this study, we did not encounter such a problem, and therefore did not adopt non-linear combinations of the outputs.

CapsNet as a Mixture of Experts: Now, we revisit the idea of the capsule networks and show how they can be viewed within the mixture of experts framework. In other words, we show that a CapsNet is a series of consecutive MoE layers such that each lower level capsule with instantiation vector $\boldsymbol{u}_i$ serves as an expert to predict the output of the capsule in the next layer with instantiation vector $\boldsymbol{s}_j$.

Each capsule (among N_{PrC} number of primary capsules) with instantiation vector $\boldsymbol{u}_i$, for $(1 \leq i \leq N_{PrC})$, makes predictions $\hat{\boldsymbol{u}}_{j|i}$, through Eq. (2.7). Consequently, each capsule (among N_{PaC} number of parent capsules) with instantiation vector $\boldsymbol{s}_j$, for $(1 \leq i \leq N_{PaC})$, receives predictions from all the lower level primary capsules. Each primary Capsule i, therefore, can be considered as an expert making predictions for all the parent (final) capsules. Contribution of each capsule expert i to each final capsule j is represented by c_{ij}, which is basically similar to g_i in an MoE framework, with the difference that in the conventional MoE formulation, each expert contributes equally to all the outputs, whereas capsule experts have different contributions to different final capsules. This is the reason why the notation of c_{ij} is used instead of c_i. The instantiation parameter of each final Capsule j is calculated according to Eq. (2.11) incorporating predictions from all the experts. Another difference between capsule experts and conventional MoE ones is that the gating model in the latter case is typically a simple or advanced machine learning model, whereas in the former case, routing by agreement serves as the gate to determine contribution through Eq. (2.8) to (2.11). It is also worth noting that Eq. (2.10) ensures that contributions to each final capsule j sum to one, satisfying the requirement of an MoE approach as in Eq. (3.12).

Having the aforementioned discussion in mind, each CapsNet itself is a series of mixtures of experts. In the proposed MIXCAPS, the CapsNets themselves are utilized as single experts. Therefore. MIXCAPS can be considered as a hierarchical MoE technique. It is also interesting to study how the calculation of c_{ij}s resembles the calculation of experts' weights in an MoE approach. Generally speaking, there are several solutions to an MoE problem [118]. An Expectation Maximization (EM) algorithm is one applicable solution, through which the experts' weights are considered as hidden variables, whose posteriors are estimated in the E-step, as follows

$$p(z_i^n|\boldsymbol{t}^n, \boldsymbol{x}^n) = \frac{p(\boldsymbol{t}^n|z_i^n = 1, \boldsymbol{x}^n)p(z_i^n = 1|\boldsymbol{x}^n)}{p(\boldsymbol{t}^n|\boldsymbol{x}^n)}, \tag{3.18}$$

where binary variable z_i^n is one when instance n is assigned to expert i, and zero otherwise. Term $p(z_i^n|\boldsymbol{t}^n, \boldsymbol{x}^n)$ represents the posterior probability of z_i^n given input vector $\boldsymbol{x}^n$ and target vector $\boldsymbol{t}^n$. Following the Bayes' rule, this posterior is calculated using the likelihood term $p(\boldsymbol{t}^n|z_i^n = 1, \boldsymbol{x}^n)$ and the prior over z_i^n, denoted by $p(z_i^n = 1|\boldsymbol{x}^n)$. All the terms appearing in Eq. (3.18) can be calculated through the MIXCAPS framework. The likelihood term can be replaced by the output of the expert capsule networks $o_i^{n(1)}$, which denotes the probability of malignancy for Instance n, based on the i^{th} expert. The prior probability can also be estimated using the output of the gating model g_i^n denoting the probability of assigning Instance n to Expert i. The posterior, therefore, can be defined as

$$p(z_i^n|\boldsymbol{t}^n, \boldsymbol{x}^n) = \frac{g_i^n o_i^{n(1)}}{\sum_j^M g_j^n o_j^{n(1)}}, \tag{3.19}$$

where M is the number of experts. The part-whole relationships in a capsule network and how the routing process resembles the MoE framework is also recently investigated in Reference [119], where the iterative procedure is replaced with a self-routing mechanism. This Reference, however, does not consider a hierarchical mixture of experts by using the capsule networks as individual experts.

To further shed light on the MoE view of CapsNets, it would be interesting to note that the EM formulation of the MoE closely resembles the weight update process of a multiple model (MM) [120] approach. In MM formulation, observations are sequentially generated from different models and the goal is to identify the contribution of each single model i given all the observations up to the current time ($\boldsymbol{Y}^k$), as follows

$$p(z_i^k|\boldsymbol{Y}^k) = \frac{p(\boldsymbol{y}^k|z_i^k = 1, \boldsymbol{Y}^{k-1})p(z_i^k = 1|\boldsymbol{Y}^{k-1})}{\sum_{j=1}^M p(\boldsymbol{y}^k|z_j^k = 1, \boldsymbol{Y}^{k-1})p(z_j^k = 1|\boldsymbol{Y}^{k-1})}, \tag{3.20}$$

where $\boldsymbol{y}^k$ is the most recent observation. Comparing Eq. (3.20) with Eq. (3.19), it can be seen that while the prior in an MoE approach is determined based on the current input vector, it is calculated based on the previous observations, in the MM case. In other words, in MM, the prior is iteratively replaced with the posterior. The updates of coefficients in the routing by agreement process of the CapsNet is similar to the weight updates in MM. In particular, in each round of the routing by agreement, the previously calculated c_{ij} serves as the prior to compute the coefficient in the next round.

Table 3.8: Performance of the proposed MIXCAPS compared to that of a single capsule network, a mixture of CNNs and a single CNN. Numbers in parenthesis show the 95% confidence intervals.

Model	**Sensitivity**	**Specificity**	**Accuracy**	**AUC**
Proposed MIXCAPS	**89.5**(89.3, 89.7)%	**93.4**(93.2, 93.6)%	**90.7**(90.6, 90.8)%	**0.956**(0.955, 0.956)
Single capsule network	86.1(85.7, 86.4)%	90.8(90.5, 91.1)%	88.6(88.5, 88.7)%	0.938(0.937, 0.939)
Mixture of CNNs	87.5(87.1, 87.8)%	91.3(91.1, 91.6)%	89.5(89.4, 89.7)%	0.948(0.946, 0.948)
Single CNN	84.1(83.7, 84.5)%	88.3(88, 88.7)%	88.3(88.2, 88.4)%	0.937(0.935, 0.938)

Table 3.9: Performance of the proposed MIXCAPS compared to that of an ensemble of capsule networks. Numbers in parenthesis show the 95% confidence intervals.

Model	**Sensitivity**	**Specificity**	**Accuracy**	**AUC**
Proposed MIXCAPS	**89.5**(89.3, 89.7)%	**93.4**(93.2, 93.6)%	**90.7**(90.6, 90.8)%	**0.956**(0.955, 0.956)
Ensemble of capsule networks	86.3(86, 86.6)%	92.5(92.2, 92.8)%	89.6(89.5, 89.8)%	0.95(0.949, 0.951)

Experiments on Lung Dataset

Three different experiments on lung cancer malignancy prediction are presented. The main objective is to evaluate performance of the proposed MIXCAPS framework and compare its capabilities with those of its state-of-the-art counterparts. Results are obtained with 200 iterations of bootstrapping, where in each iteration, 80% of the data is sampled (with replacement) from the whole dataset. 20 % of the training dataset is then randomly extracted for validation. A 95% confidence interval (CI) is calculated for all the performance metrics. Adam optimizer with 10 epochs and batch size of 16 is used for training.

Experiment 1: Our first experiment is to compare the performance of the proposed MIXCAPS with a single capsule network, a mixture of CNNs, and a single CNN, as shown in Table 3.8, where performance is measured in terms of sensitivity, specificity, accuracy, and area under the curve (AUC). The architecture of the single capsule network is exactly the same as the CapsNet experts. We tried to keep the complexity as similar as possible to the MIXCAPS, when designing the mixture of CNNs. In particular, the gating network exactly resembles that of the MIXCAPS. The CNN experts consist of two convolutional layers with 256 filters, similar to the experts in the MIXCAPS. The convolutional layers are followed by a dense layer with 32 neurons (the same as the dimension of the last capsule layers), and the final softmax layer for nodule malignancy prediction. The single CNN has also the same architecture as the experts in the mixture of CNNs. As shown in Table 3.8, MIXCAPS outperforms its three aforementioned counterparts, in terms of sensitivity, specificity, accuracy, and AUC. In particular, to gain an insight on what components of the proposed MIXCAPS lead to its superior performance, it is worth noting the higher performance of a single capsule network compared with a single CNN, showing the advantage of a capsule network-based design. Furthermore, the superiority of the proposed MIXCAPS over the single capsule network, as well as the mixture of CNNs over the single CNN, illustrates the benefit of a mixture of experts framework.

Experiment 2: In a second experiment, we compare the proposed MIXCAPS with an ensemble of two capsule networks. The goal of this experiment is to investigate whether the trainable gating network is an effective component of the MIXCAPS by replacing it with a non-trainable average voting of the two capsule networks. The utilized ensemble model consists of two capsule networks with the same architecture as the experts in the MIXCAPS. However, the gating model is removed, and the final output is the average of the outputs of the two capsule networks. This averaging is performed through both training and testing steps. In other words, Eq. (3.13), which calculates the final output, is modified as follows

$$\boldsymbol{o} = 0.5 \times \boldsymbol{o}_1 + 0.5 \times \boldsymbol{o}_2. \tag{3.21}$$

Table 3.9 shows the obtained results according to which, we can observe that replacing the gating model with an average voting degrades the performance. It is worth mentioning that ensemble of capsule networks, although having lower performance compared to the MIXCAPS, outperformed the stand-alone capsule network architecture illustrating the benefits of ensembling capsule networks.

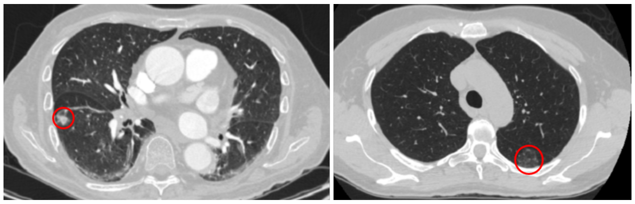
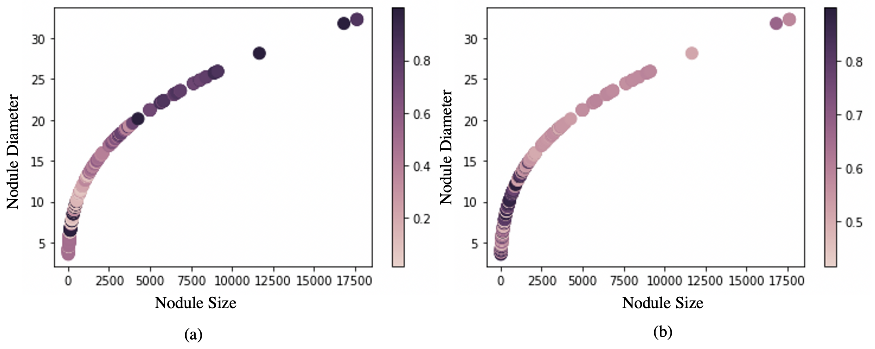


Figure 3.18: Example of nodules assigned to experts based on their volume and diameter. The nodule on the left, which has a lower probability of belonging to the first expert, is smaller in terms of volume and diameter compared to the nodule on the right.



(a)

(b)

Figure 3.19: Activation magnitude of the true class capsule, based on the tumor size and diameter. The darker the color, the higher the activation. (a) The true class capsule of the first expert has higher activation for larger nodules. (b) The true class capsule of the second expert has higher activation for smaller nodules.

Experiment 3: We begin this experiment by measuring the performance of the individual experts on the test set. In other words, after the MIXCAPS is trained on the training set, we use experts separately for the final decision on the test set. Doing so, we observed that the accuracy of the first and the second experts decreased to 66% and 60%, respectively. This implies that the performance of the MIXCAPS is related to leveraging the capabilities of both of the experts. Next, we conduct an experiment to gain an insight on how the data instances are split between the two experts. The LIDC-IDRI dataset is accompanied by a few nodule-related properties, determined by the radiologists. These features include volume, diameter, x center of mass and y center of mass. We calculated the correlation between the output of the gating network and these features. While the correlations with volume and diameter are 0.58 and 0.77, respectively, we observed no correlation with the centers of mass. It should be noted that the inputs to the proposed MIXCAPS are cropped nodule regions. In other words, the model has no access to the location of the nodule. Therefore, the almost zero correlations with the centers of the mass is completely expected. The observed correlations between the gate outputs and the volume and diameter imply that larger nodules have higher probabilities of being assigned to the first expert. Fig. 3.18 shows two nodules in the test set. The left nodule, which has a volume of 496.32 and diameter of 9.823, has a low probability of belonging to the first expert, whereas the nodule on the right, with a volume of 6663.44 and diameter of 23.347, has a high probability of being assigned to the first expert. In other words, the first expert tends to handle larger nodules, compared to the second expert. Finally, it is worth exploring how the individual experts respond to the nodules assigned to them. To this goal, we measured the activation of the capsule, related to the true class, for both of the experts, where activation refers to the length of the capsule. Consequently, we studied the relation between this activation and the size and diameter of the nodule, as shown in Fig. 3.19. According to this figure, the first expert (plotted on the left hand side) has higher activation for larger nodules, whereas the second expert (plotted on the right) has higher activation for smaller nodules. This observation is also consistent with how the gating model assigns the nodules. While the gating model tends to assign the larger nodules to the first expert, the true class capsules in the first expert have high activation for these nodules. The same rationale holds for the second expert being assigned with smaller nodules.

Table 3.10: Performance of the proposed MIXCAPS with BoxCaps as experts. Numbers in parenthesis show the 95% confidence intervals.

	MIXCAPS-BoxCaps	**BoxCaps**
Accuracy	**91.3** (91.1, 91.5) %	90.9 (90.2, 91.5) %
Sensitivity for Meningioma	77.5 (77.1, 77.9) %	**80.1** (76.2, 84) %
Sensitivity for Glioma	**95.9** (93.2, 98.5) %	92 (90, 94.1) %
Sensitivity for Pituitary	**97.7** (97.2, 98.3) %	97.2 (95.6, 98.9) %
Specificity for Meningioma	**96.1** (96, 96.1) %	94.1 (92.7, 95.5) %
Specificity for Glioma	88.7 (87.6, 89.8) %	**89.8** (88.4, 91.2) %
Specificity for Pituitary	**88.7** (86.2, 91.2) %	88.1 (86.9, 89.3) %

Although MoE techniques are shown to be able to improve the classification performance, they typically face an objection related to the high computational cost at the test time. This problem, however, can be dealt with by using distillation [121]. Therefore, in our future studies, we will focus on distilling MIXCAPS into a smaller and more time-efficient model.

Experiments on Brain Dataset

To investigate whether the MIXCAPS can be generalized to brain tumor classification, we replaced the capsule experts in MIXCAPS with the previously designed BoxCaps architecture, as shown in Fig. 3.20. We then tested the resulting framework on the brain tumor dataset [43], where train, validation, and test splits are obtained from the same bootstrapping approach used for the LIDC-IDRI dataset. Table 3.10 presents the obtained results, according to which, the MoE approach leads to higher accuracy compared to a single BoxCaps. Furthermore, the MoE approach leads to higher sensitivity for Glioma and Pituitary, and higher specificity for Meningioma and pituitary tumor types.

To investigate whether the individual experts specialize on certain tumor types, after training the MIXCAPS-BoxCaps, we evaluated the experts separately on the test set. We also calculated what percentage of each tumor type was assigned to the experts by the gating model. Consequently, we observed that while 94% of the Gliomas were assigned to the first expert, this expert had a high sensitivity of 99% for the underlying tumor type. This expert, however, had very low sensitivities for the other two tumor types. The second expert, on the other hand, received 88.6% of the Meningiomas and Pituitaries, leading to 95% and 96.2% sensitivities, respectively.

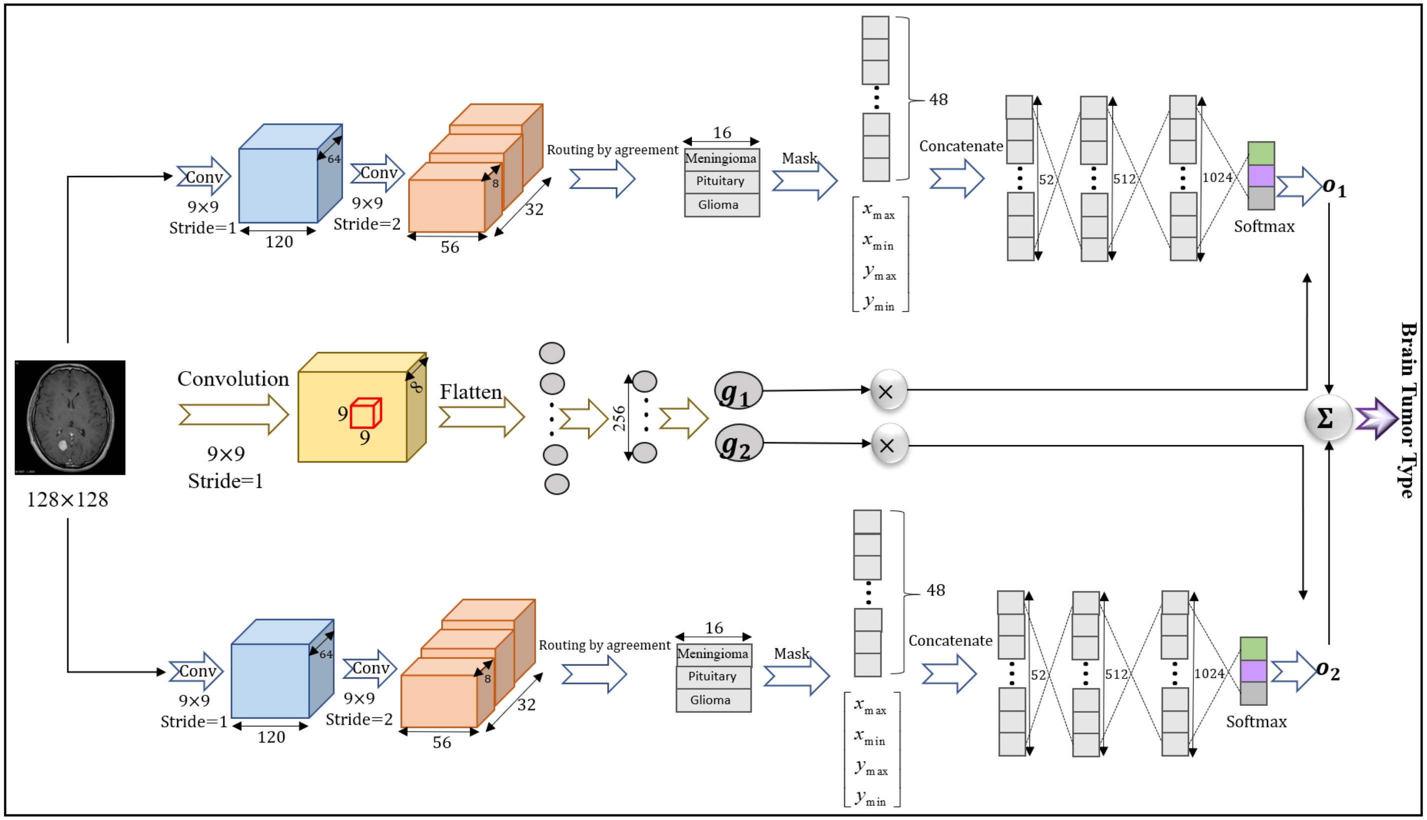


Figure 3.20: MIXCAPS architecture with BoxCaps as experts for brain tumor type classification.

Finally, we conduct another experiment to study if the provided boundary box is the only important factor leading to the obtained result. In other words we need to make sure that the input images are not ignored by the model, simply because the boundary box itself can determine the tumor type. To this end, we gradually added zero-mean Gaussian noise to input images and calculated the model's accuracy. It is observed that while a noise with a standard deviation (STD) of 0.01 does not change the accuracy, increasing STD to 0.1 and 0.5 degrades the accuracy to 84.44% and 76%, respectively. This experiment shows that while the boundary box assists the classification, it does not replace the input images.

3.3 Conclusion

In this Chapter, we targetted the problem of tumor type classification using deep learning-based radiomics by focusing on brain tumor classification in Section 3.1 and lung tumor classification in Section 3.2. For the former category, we have presented a CapsNet architecture, referred to as BoxCaps, that incorporates both the raw MRI brain images and the tumor coarse boundaries in order to classify the tumors. The proposed CapsNet architecture has two main advantageous: (i) First, the need for tumor exact annotation is eliminated, and; (ii) Second, it helps the CapsNet to focus on the main area, and at the same time, consider its relation with surrounding tissues. Our results show that the proposed approach is capable of increasing the classification accuracy, compared to the previous CapsNets and CNN architectures. Capitalizing on the success of ensemble techniques in different domains, we improved BoxCaps by incorporating a boosting approach. Since, similar to most of the deep learning models, CapsNet does not capture model uncertainty in its predictions, we equipped CapsNet with a Bayesian framework to not only model the posterior distributions over the weights, but also estimate the mean prediction and entropy, having the benefit of keeping human in the inference loop. For the problem of lung nodoule type classification, i.e., to address the lung nodule diagnosis problem, we proposed a 3D multi-scale capsule network, capable of distinguishing between benign and malignant lung nodules. This model, which benefits from 3D inputs from three different scales, can capture local and global features from the tumor. Our experiments show that the proposed model outperforms its counterparts, from different aspects such as accuracy and sensitivity. Finally, we concluded this chapter by proposing a CapsNet-based mixture of experts, having the promise of assigning similar instance to separate experts, utilizing a trainable gating network.

Chapter 4

Time-to-Event Outcome Prediction

In Chapter 1, three application domains for deep learning-based radimoics, i.e., tumor classification, time-to-event outcome prediction, and COVID-19 diagnosis, have been identified. In Chapter 3, we focused on the first application. In this chapter, we focus on the second task to predict time-to-event outcome for lung cancer patients. The reminder of the chapter is organized as follows: Section 4.1 describes the intuition and problem statement, followed by the proposed DRTOP framework in Section 4.2. Results and experiments are presented in Section 4.3. Finally, Section 4.4 concludes the chapter.

4.1 Introduction

Significant recent progress in the biological understanding and tumor heterogeneity of non-small cell lung cancer calls for treatment individualization. Specific clinical endpoints are used in clinical trials to measure the clinical benefit of a specific treatment [122, 123]. Although overall survival (OS) remains the gold standard, other clinical endpoints such as recurrence free survival (RFS), distant control (DC), and local control (LC) measure different and significant aspects of the clinical benefit of treatment. Inherent difficulties to assess these clinical outcomes such as the lengthy duration of the follow-up needed until the time of event and the various parameters, unrelated to the primary cancer, affecting the result during follow-up, have led to a surge for developing surrogates that can predict clinical outcomes noninvasively. Recently, radiomics, which is the process of extracting high throughput quantitative and semi-quantitative features from medical images aiming at diagnosis, classification or prediction of outcomes, has attracted much attention, showing promising results [1,20,24,124–131].

Studies, investigating the relation between radiomics and time-to-event outcomes (e.g., survival and/or recurrence), have mostly focused on hand-crafted radiomics, referring to extracting pre-defined features. Using pre-treatment Computed Tomography (CT) images, Sun *et al.* [132] have extracted 339 pre-defined features from the segmented lung tumor volume, to predict the patients' OS. These features, including shape, size, texture, and intensity statistics, are shown to be predictors of the OS, when going through a set of feature selection and machine learning methods. The prognostic value of hand-crafted radiomics features for OS in lung cancer is also studied by Timmeren *et al.* [133], where CT-based extracted features led to a concordance index (a measure of model accuracy) of 69%. Khorrami *et al.* [134], recently, investigated the correlation of CT-based features with OS and time to progression (TTP) in lung cancer patients treated with chemotherapy, and found a high predictive ability for the extracted features. Although hand-crafted radiomics has shown correlation between imaging modalities and the clinical outcomes, its practical application is limited by the fact that features are pre-defined. Furthermore, hand-crafted radiomics requires the exact segmented region of interest (ROI), being highly dependent on the quality of the segmentation. Obtaining an accurate segmentation is burdensome and subject to inter-observer variability [135], challenging the reliability of the result.

Considering the potential of radiomics, and at the same time, the limitations associated with hand-crafted radiomics, there has been a recent surge of interest [1, 25, 52, 108, 136] in using deep learning, especially Convolutional Neural Networks (CNNs) [137,167], to extract radiomics features. In deep learning-based radiomics, features are not pre-defined, and do not require the segmented ROI. Therefore, the model can be trained in an end-to-end fashion, with the goal of improving the overall prediction accuracy. Zhu *et al.* [139] developed a CNN to predict OS in lung cancer patients and trained the model on pathological images of the lung tumor, leading to a 63% concordance index.

Most of the studies on deep learning-based time-to-event outcome prediction in cancer have focused on features extracted from CT images, which capture only

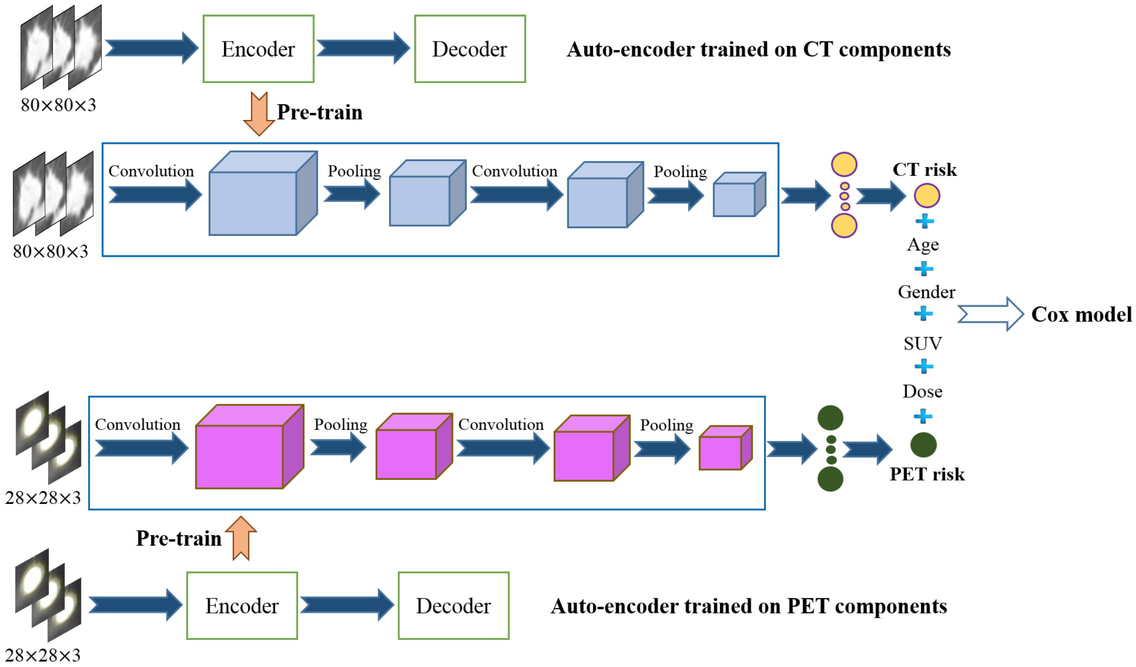


Figure 4.1: Proposed DRTOP model, where 3D CT and PET images go through separate networks, which are unsupervisedly pre-trained on an independent dataset. The outputs of the two networks (referred to as the CT risk and the PET risk) are combined with other clinical factors and fed to a Cox PHM.

anatomical information. 18-Fluorodeoxyglucose Positron Emission Tomography/Computed Tomography (FDG PET/CT), which combines anatomic data with functional and metabolic information, is the standard of care and has become an integral part of lung cancer staging in clinical practice [140]. The focus of this chapter is to propose a novel deep architecture, referred to as DRTOP [8], based on staging PET/CT images to predict pre-defined clinical endpoints in a cohort of lung cancer patients before the initiation of treatment. The schematic of our proposed DRTOP model is shown in Fig. 4.1.

4.2 DRTOP Architecture based on Staging PET/CT Images

Data Description and Pre-processing: The in-house dataset we used in this study consists of 132 lung cancer patients (65 women, 67 men) with an average age of 74.65 (range: 52-92), having 150 lung tumors in total, (treated between April 2008

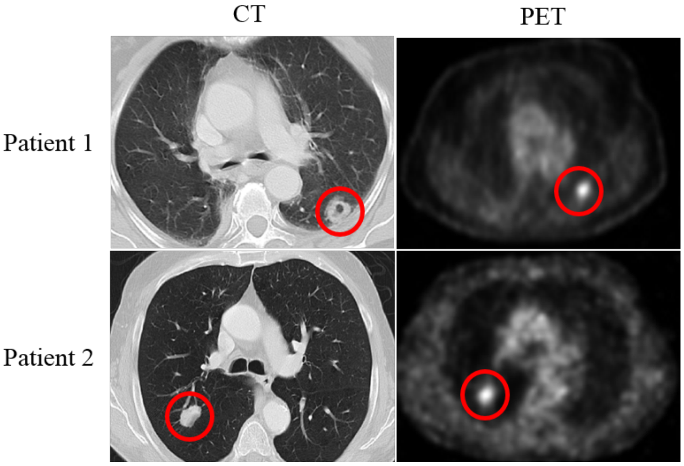


Figure 4.2: The CT and PET components of the PET/CT, for patient 1 and patient 2, show the tumors in the superior segment of the left lower lobe and the right lower lobe respectively. CT and PET images for each patient are captured at the same level.

and September 2012), who underwent staging pre-treatment PET/CT. Tumors visible in both CT and PET components are annotated by a thoracic radiologist, with 18 years of experience in thoracic imaging (A. O.), using an in-house software described in Reference [19] as follows: Each lesion was contoured on every sequential slice that was visible on CT as increased homogeneous or ground glass density compared to surrounding normal lung parenchyma. Attention was made so that volume averaging areas, and adjacent vascular structures were not included in the regions of interest. The segmentation/contouring of the lesions on the PET images was performed manually on all the sequential images showing increased FDG uptake in the corresponding area of the tumor, which was either the same area covered on the equivalent CT images or slightly smaller. Fig. 4.2 shows the observed tumor for two patients on the CT and PET component, at the same level. Other characteristics that were entered and assessed in the analysis include age, gender, SUVmax, and radiation dose (prescribed biological effective dose). All the patients had early stage lung cancer (N0M0) and were treated with a specific high dose and focused radiotherapy method (SBRT) [19]. Post-treatment patients were followed up for a median period of 27 months, during which different observations, including local recurrence, regional recurrence, lobar recurrence, distant recurrence, and death were recorded. In this work, we have focused on four different outcomes: (i) Overall survival (OS), which is defined as the time from the SBRT to the date of the death or final follow-up visit; (ii) Recurrence free survival (RFS), referring to the time from SBRT to the earliest of recurrence (local, lobar, regional, or distant), second cancer, death or final follow-up visit; (iii) Local control (LC), defined as the absence of local (within the area of the planning target volume) recurrence, and; (iv) Distant control (DC), calculated as the absence of recurrence outside of local, lobar or regional recurrences. There are patients who have more than one lung tumor. Since the outcomes of OS, DC, and RFS are related to the patient and not to each single tumor, we decided to take the tumor with the highest SUV. However, LC is tumor-related, and therefore, all the 150 tumors are treated as data instances.

All the images are cropped based on the annotations provided by our experienced Radiologist to only contain the tumor region. As the proposed DRTOP model requires fixed size inputs and tumors have different sizes, we have zero-padded the cropped tumor regions. In other words, cropped tumors are placed in the middle of a black image (intensity of zero), whose size is determined based on the largest tumor available in the dataset. The largest tumor available is of size 80×80 pixels in CT images and 28×28 pixels in PET images. All the images are, therefore, first cropped to completely fit the tumor. Then, cropped CT scans are placed in 80×80 black images, whereas cropped PET scans are placed in 28×28 black images. Determining the size of the inputs, based on the largest tumor, to ensure all the target area is covered, is a common practice in deep learning-based cancer image analysis [142].As the inputs to our model are 3D images, where the third dimension is of size 3, three cropped slices, for each tumor, are stacked together. The middle image is the tumor middle slice, and the other two are the two immediate neighbors of the middle slice. At the end, each patient/tumor is associated with two 3D inputs, one generated from the PET component, and the other generated from the CT component.

Here we further elaborate on the choice of a 3-channel input. As shown in Fig. 4.3, number of the tumor-containing slices, in our in-house dataset, significantly varies from one patient to another (between 3 and 42). The proposed DRTOP architecture requires a fixed-size input. This means that in case of selecting a higher number of channels, all inputs having less number of tumor-containing slices, have to be accompanied with healthy slices, in order to maintain a fixed size. Accordingly, selecting a higher number of channels leads to the following two important challenges: (i) First, it requires advanced memory resources, and; (ii) Second, it makes some

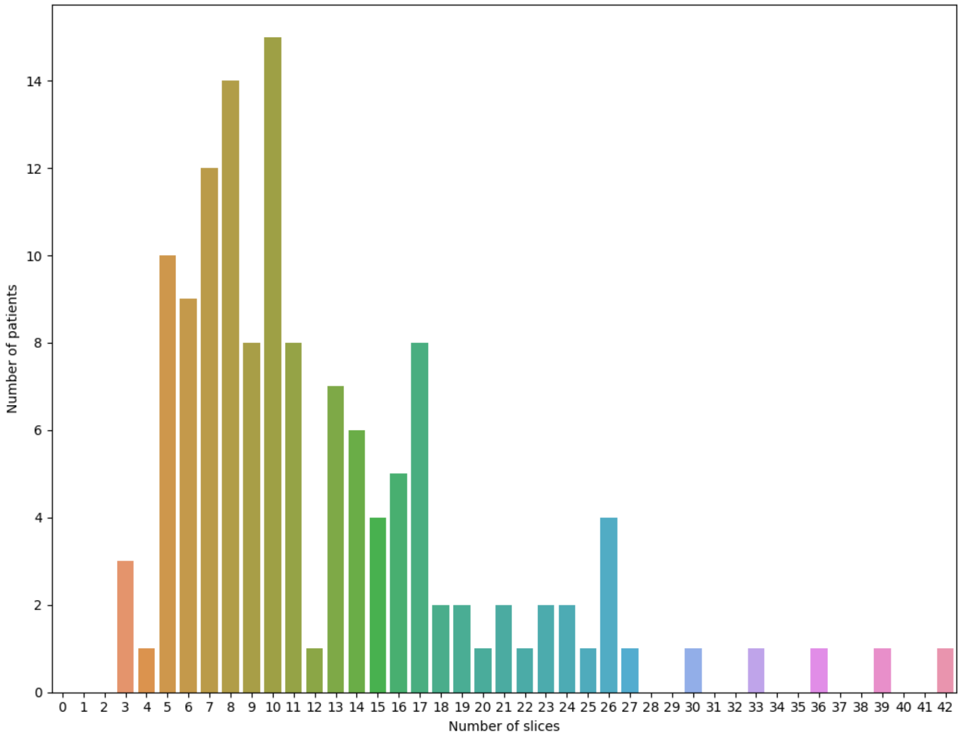


Figure 4.3: Frequency of number of tumor-containing slices, that can vary from 3 to 42.

tumors too small to be distinguished from surrounding tissues [143]. Furthermore, the 3-channel input has been previously investigated in several studies, leading to satisfying results. For instance, in Reference [144], 3-channel CT scans are used to predict short and long-term survival in lung cancer, using CNNs, and it has been shown that the 3-channel input outperforms the single channel one. The 3-channel input is also utilized in References [145] and [146], for classifying breast tumor and mediastinal lymph node metastasis of lung cancer, respectively, using CNNs.

In order to validate our model and also the hand-crafted method, we have split the dataset into two independent set of training (80%) and testing (20%) instances. The training dataset is used to train our proposed model, and also the hand-crafted method, whereas the test set remains unseen during the training, and is used at the end for evaluating the models.

It is worth mentioning that for lung cancer survival analysis, large datasets are

Table 4.1: Datasets used in the literature for lung cancer time-to-event outcome predictions.

Reference	**Number of patients**	**Difference with our dataset**	**Availability**
Wu *et al.* [124]	101	Only PET images are utilized and outcome is distant metastasis.	Not public
Pyka *et al.* [125]	45	-	Not public
Huang *et al.* [126]	282	Only CT images are utilized and outcome is disease-free survival.	Not public
Sun *et al.* [132]	422	Only CT images are utilized and outcome is overall survival.	Public [20]
Khorrami *et al.* [134]	125	Only CT images are utilized and outcome is overall survival and time to progression.	Not public
Wang *et al.* [160]	129	Only CT images are utilized and outcome is overall survival.	Not public
Xu *et al.* [141]	179	Only CT images are utilized and patients are treated with chemoradiation.	Not public

scarce and very difficult to acquire, as patients need to be followed up for years. Studies investigating the problem of lung cancer time-to-event outcome prediction, a few of which are listed in Table 4.1, therefore, evaluate their models on relatively small datasets. In order to evaluate the proposed DRTOP model, dataset needs to include both PET and CT images that are contoured by an expert, which limits us to the in-house dataset with 132 patients. Furthermore, outputs of the model, i.e., OS, RFS, LC, and DC, are required to be available for all the patients in the dataset. As we have shown in Table 4.1, the dataset used in Reference [125] is the only one that includes all the DRTOP's requirements. This dataset, however, is not publicly available and is limited to 45 patients. To the best of our knowledge, the NSCLC-Radiomics dataset [20] is the only publicly available data that focuses on the lung cancer survival analysis. Nevertheless, it is accompanied with only CT images and the outcome is limited to the overall survival.

CNN Architecture of the DRTOP Model: The CNN architecture we used in this work is shown in Fig. 4.1. We have adopted two separate networks, for CT and PET components, each of which contains two convolutional layers (with 3×3 filters, 32 feature maps, and rectified linear units), two pooling layers (with 2×2 filter size) and two fully connected (FC) layers. The first and second FC layers contain 32 and 1 neurons, respectively. While the first FC utilizes rectified linear units, the second one has a linear activation. Both CNNs are trained separately with the goal of maximizing the Cox partial likelihood. The optimization method is a stochastic gradient descent (SGD), with a learning rate of 10^{-5}. Number of epochs is set to $2,000$, and while most of the studies on deep learning-based time-to-event analysis feed the model with the whole dataset at once, we used a batch size of 32 [147], to prevent the network from over-fitting the training set. The outputs of the last fully connected layers are treated as radiomics signature (risk), and fed to a Cox PHM, along with the other clinical factors (age, gender, SUV, and radiation dose).

One problem associated with CNNs is that they, typically, require large datasets to be able to learn a useful mapping from the input to the output. Otherwise, the network over-fits the training set, leading to poor predictions for the test set. Large dataset is, however, difficult to collect in the medical field. One solution to compensate for the lack of large dataset is to pre-train the model with, preferably, a similar dataset [148]. Pre-training helps the network with learning the data distribution. Consequently, when training the model supervisedly on the main dataset, weights are initialized by the pre-trained values instead of the random ones, getting one step closer to the optimal solution. The Convolutional Auto-encoder (CAE) [149,150], we adopted in this work for the pre-training, is explained next.

Pre-training with Convolutional Auto-encoders: Auto-encoders are unsupervised neural networks that are only fed with the input, without any additional information or labels. The network is aimed to learn features from the input that are useful in reconstructing the input. Auto-encoders consist of two main components, i.e., the encoder, which learns features from the input, and; the decoder, which uses learned features to reconstruct the input. The CAEs are variants of the original Auto-encoders with embedded Convolutional layers, making them powerful models for unsupervised training of the image inputs. In this work, two separate CAEs are trained on the PET and CT components, where the encoder's architecture is exactly the same as the main CNN architecture described in the previous sub-section. The

CNNs are, consequently, initialized with the weights learned in the CAEs, through the unsupervised training.

The dataset we used for the unsupervised pre-training is different from the main dataset. However, it includes pre-treatment PET/CT images of 86 lung cancer patients from a previous work [151]. This dataset does not contain the time-to-event outcomes. Images in this dataset are also annotated by our thoracic radiologist (A.O.), and pre-processed using exactly the same approach as the one used for pre-processing the main dataset.

Cox Proportional Hazards Model (PHM):

The DRTOP model was trained, separately, for all four outcomes, and calculated the CT and PET risks. These two risks, along with four clinical factors are entered into the Cox PHM, using a stepwise selection of the variables. In other words, the final model includes only the significant predictors, where significance is evaluated based on an F-test of the obtained coefficients. Therefore, to assess the significance of a coefficient, the Cox PHM is trained after excluding the underlying variable (restricted model), and compared against training the model, including all the variables (unrestricted model). The significance level is set to 0.05, and only the variables associated with p-values less than the significance level remain in the model. Table 4.2 shows the four time-to-event outcomes along with their significant predictors. Hazard ratio (HR) measures the effect of the predictors on the outcome. Concordance index [152], presenting the quality of ranking, is also computed for all the four outcomes. The PHM formulation, based on our predictors is as follows

$$h(t|x_i) = h_0(t)\exp^{\left(\beta_1 \times CT_i \ + \ \beta_2 \times PET_i \ + \ \beta_3 \times Age_i \ + \ \beta_4 \times Gender_i \ + \ \beta_5 \times SUV_i \ + \ \beta_6 \times Dose_i\right)} \tag{4.1}$$

where $h(t|x_i)$ refers to the hazard at time t for the i^{th} patient. Term $h_0(t)$ is the base-line hazard, and β_is ($1 \leq i \leq 6$) are the coefficients (covariates) to be learned with the objective of maximizing the Cox partial likelihood.

It is worth mentioning that in design of the proposed DRTOP model, we have chosen to use the final deep learning output as the inputs to the Cox PHM. The rationale behind this design is to prevent the 64 features, extracted from the layer before the final one, from dominating the Cox PHM, and cancel out potential effects of the clinical factors (age, gender, SUV, and radiation dose). The incorporated strategy is similar in nature to the approach adopted in Reference [134], where a

Table 4.2: Results from the proposed DRTOP model. HR stands for hazard ratio (exponent of the obtained coefficient). Significant predictors are obtained based on an F-test, with a significance level of 0.05. Concordance index is calculated on the test set.

Clinical outcome	**Significant predictors**	**Concordance index**
OS	CT risk (HR: 1.35, p-value: < 0.005), PET risk (HR: 0.67, p-value: < 0.005), Age (HR: 1.01, p-value: 0.02)	68%
RFS	PET risk (HR: 1.18, p-value: < 0.005), SUV (HR: 1.13, p-value: < 0.005)	40%
DC	CT risk (HR: 1.06, p-value: < 0.005), SUV (HR: 1.09, p-value: 0.02)	63%
LC	CT risk (HR: 2.66, p-value: 0.03)	37.5%

Table 4.3: Hand-crafted features, extracted in DRTOP study.

Category	**Description**	**Sub-category**
First Order Radiomics	Distribution of pixel intensities and ROI shape.	
• Shape Features	Quantify the geometric shape of the tumor [17]	Area regularity (1), Perimeter regularity (2), Region bilateral symmetry (4).
• Intensity Features	Derived from a single histogram [17].	Size of the tumor (number of pixels), Mean gray level, Standard deviation, Median gray level, Minimum pixel intensity, Maximum pixel intensity, Kurtosis, Skewness [17, 153].
Second Order Radiomics (Texture Features)	Relations between pixels to model intra-tumor heterogeneity [17].	Contrast, Energy, Correlation, Homogeneity, Entropy, Normalized Entropy.

Least Absolute Shrinkage and Selection Operator (LASSO) Cox model is, first, used to extract the most important radiomics features, before going through the final Cox PHM, along with other clinical factors.

Random Survival Forest (RSF): An RSF model [154] is a collection of several survival trees, each of which is constructed using a randomly drawn sample of the data and underlying variables. Each survival tree is separately trained, based on a logrank splitting rule, which tries to maximize the survival difference between the daughter nodes. While each tree outputs a separate CHF for each patient, the final outcome is the ensemble CHF. The Nelson-Aalen estimator is used to calculate the CHF, denoted by $\hat{H}$, at each terminal node h, and is given by

$$\hat{H}_h(t) = \sum_{t_{l,h} \leq t} \frac{d_{l,h}}{Y_{l,h}}, \tag{4.2}$$

where $t_{l,h}$ denotes a distinct event time at node h, and $d_{l,h}$ and $Y_{l,h}$ are number of death and patients at risk, respectively, at time $t_{l,h}$. In this study, the RSF model consists of $10,000$ survival trees. The maximum depth is set to 10, and the minimum node size is 10. To obtain the important predictors, a variable importance (VIMP) [154] approach is adopted. Based on this approach, for each variable, the prediction error is calculated for the original RSF, and also an RSF with random assignment, when encountering the underlying variable. The VIMP is then calculated as the difference between these two errors. A large positive VIMP indicates a high predictive ability, whereas a zero or negative one means no prognostic value.

Hand-crafted Radiomics: To compare the ability of our proposed DRTOP model in predicting the time-to-event outcomes, in lung cancer patients, with the hand-crafted radiomics, we have extracted 42 features from the CT and PET components. As shown in Table 4.3, these features include both first and second-order radiomics, where the former refers to features extracted mostly from the image histogram, and the latter refers to texture-related features. The "Sub-category" column presents the features we have extracted, where the numbers in the parenthesis indicate the number of features extracted from that specific category. All the features, consequently, go through a PCA, where a total of 18 features are extracted. These features are the inputs to a stepwise Cox PHM.

Concordance Index: Concordance index (c-index) is a measure of how well the patients are ranked based on a specific time-to-event outcome. Mathematically, it can be defined as [155]

$$c = \frac{1}{|\xi|} \sum_{T_i \ uncensored} \sum_{T_j > T_i} 1_{f(x_i) < f(x_j)}, \tag{4.3}$$

where $|\xi|$ denotes the number of possible ordered pairs, T_i and T_j are the time-to-event outcomes for Subjects i and j, respectively, and $f(x_i)$ is the predicted time for Subject i. The c-index varies between 0 and 1, where a c-index of 1 means a perfect prediction, and a c-index of 0.5 can be interpreted as a random assignment. In biomedical applications and in particular lung cancer survival analysis, a c-index close to 0.7 is considered as satisfying and acceptable [156, 157], however, interpretation of the computed c-index value depends on the dataset and the problem at hand.

4.3 Experiments and Performance of the Proposed DRTOP

Based on Table 4.2, the OS can be predicted by CT risk, PET risk, and age, where PET risk with the HR of 0.67 has a negative impact (protective effect) on the OS, CT risk with the HR of 1.35 has a positive (an increased risk) impact, and the impact of age is relatively small. The obtained HRs can be interpreted as follows: (i) one unit increase in the CT risk predictor variable results in an increase in the risk of the event occurring by 35%; (ii) Increasing the PET risk by one unit leads to a 33% risk reduction, and; (iii) One year increase of the age can only increase the hazard by 1%. The concordance index of 68% shows that the three predictors are capable of providing a satisfying ranking of the patients, with regards to the OS.

Performance of the Hand-crafted Radiomics: Out of 42 PET and CT hand-crafted radiomics features, calculated as suggested by a previous study [19], 18 features (principal components) are extracted using the principle component analysis (PCA). These features, together with SUV, age, gender, and radiation dose, are fed to a Cox PHM to explore predictive models for the specific time-to-event outcomes. Table 4.4 illustrates the obtained results. Radiomics (PC2) is the only predictor of the OS. Radiomics (PC1 and PC2), together with SUV, contribute to the prediction of RFS and DC. Neither hand-crafted radiomics nor clinical factors can predict the LC. A failed prediction for LC, using hand-crafted radiomics, does not mean a concordance index (c-index) of 0. It means that the c-index is not calculated because the Cox PHM has not found any significant predictors for the LC, where significance is assessed using an F-test. In other words, any calculated c-index, in this case, is not reliable and can be the result of a random model. The c-index is not necessarily an indicator of the predictors' performance, when they fail to statistically predict enough of the variability in the outcome. If no predictors are found for the model then the hazard function is equal to the baseline hazard. In the case of Cox PHM the baseline hazard is not estimated as it is a semi-parametric approach which was designed specifically to benefit from NOT having to estimate the baseline hazard.

Comparison of the DRTOP and Hand-crafted Radiomics: Fig. 4.4 shows the comparison between the concordance indices obtained from the hand-crafted radiomics and the proposed DRTOP model. The performance of the proposed model is

Table 4.4: Results obtained from hand-crafted radiomics. As there is no significant predictor for the LC, Concordance index is not calculated.

Clinical outcome	**Significant predictors**	**Concordance index**
OS	PC2 (HR: 0.44, p-value: 0.02)	51%
RFS	PC1 (HR: 1.57, p-value: 0.02), PC2 (HR: 0.37, p-value: < 0.005), SUV (HR: 1.14, p-value: < 0.005)	47%
DC	PC1 (HR: 1.58, p-value: 0.03), PC2 (HR: 0.33, p-value: < 0.005), SUV (HR: 1.14, p-value: < 0.005)	64%
LC	-	NA

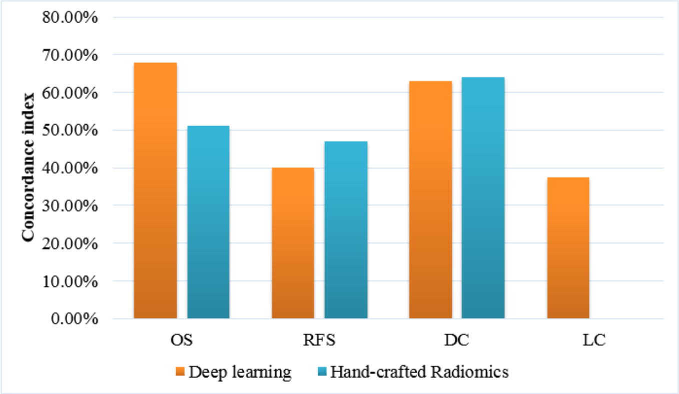
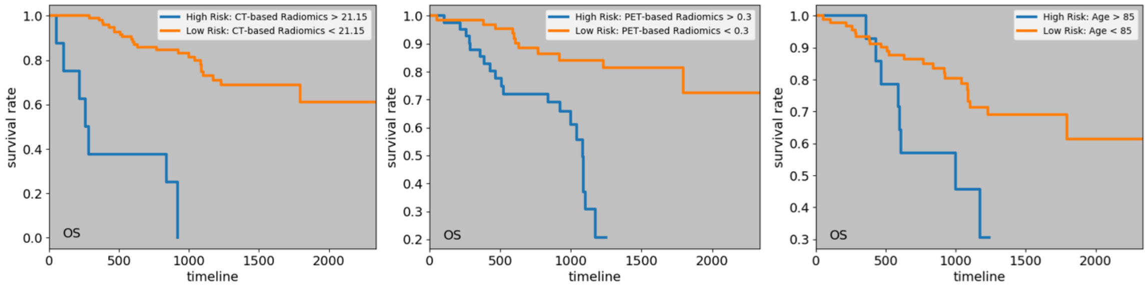


Figure 4.4: Comparison between our proposed DRTOP model and hand-crafted radiomics, based on the concordance index.

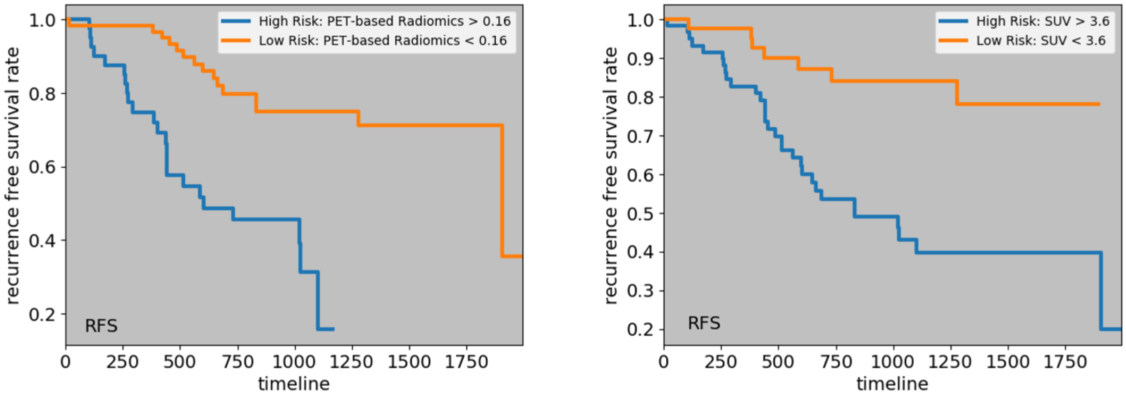
better than the hand-crafted method, in predicting the OS. Although both methods fail to provide a satisfying result for predicting the RFS, the hand-crafted radiomics has a slightly better performance. The two methods are on a par with each other, in predicting the DC, and in the case of the hand-crafted method, no significant variable is identified to predict the LC. We also attempted to predict the time-to-event outcomes, based on the combination of hand-crafted and deep learning-based features. However, this did not improve the predictive ability of the model.

Kaplan-Meier Curves and Cut-off Values: To visualize the impact of a variable on the survival function of different groups, such as low-risk and high risk ones, Kaplan-Meier estimation technique [158] is utilized. The cut-off value to identify these two groups is often calculated based on a logrank test [159], that tries to maximize the survival or recurrence difference between the two groups. Considering the significant predictors of the four time-to-event outcomes, we have computed the cut-off values, and obtained the low and high-risk groups, as shown in Figs. 4.5-4.7. The cut-off values to identify low and high-risk groups from CT risk, PET risk, and age (in years) for OS are 21.15, 0.3, and 85, respectively. In other words, a patient having a CT risk higher than 21.15, and/or a PET risk higher than 0.3, and/or age higher than 85 is considered a high-risk patient, and has a lower chance of survival compared to a low-risk patient. It should be noted that while PET risk is associated with a hazard ratio of less than one in the DRTOP model, meaning that it has a negative impact on the outcome when combined with other factors, it has a positive impact when it is the only predictor taken into account. The cut-off values obtained from



(a) (b) (c)

Figure 4.5: Kaplan-Meier curves associated with the OS, with respect to (a) CT risk, (b) PET risk, and (C) age. Cut-off values to determine the low and high-risk groups are obtained from a logrank test. All predictors, when considered independently, have positive correlations with the OS.



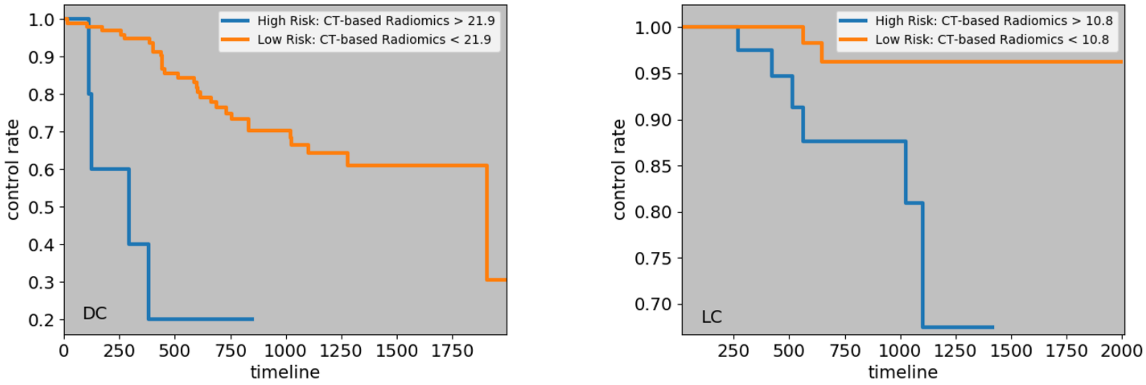
(a) (b)

Figure 4.6: Kaplan-Meier curves associated with the RFS, with respect to (a) PET risk, and (b) SUV.

the PET risk and SUV for the RFS, are 0.16 and 3.6, respectively, and the cut-off values obtained from the CT risk for DC and LC are 21.9 and 10.8, respectively.

Random Survival Forest (RSF) Analysis: Our results demonstrated that, based on an RSF model, recurrence free survival (RFS) can be predicted with a concordance index of 64%, while, based on the variable importance (VIMP) values presented in Table 4.5, all predictors, except the radiation dose, show predictive importance.

Fig. 4.8 shows one of the obtained trees from the RSF. Cumulative hazard function (CHF) is calculated for all the terminal nodes, and all the unique time points. However, only the CHF associated with the first event time is shown in this figure. It should be noted that the left terminal node is associated with a CHF of zero, meaning



(a) **(b)**

Figure 4.7: Kaplan-Meier curves corresponding to (a) DC, and (b) LC, with respect to the CT risk, where cut-off values are determined by a logrank test.

Table 4.5: Variable importance values obtained from the RSF, for recurrence free survival prediction. The negative value means no predictive importance.

Variable	CT risk	PET risk	Age	Gender	SUV	Radiation dose
VIMP	60.72	71.36	0.59	6.73	53.90	−4.82

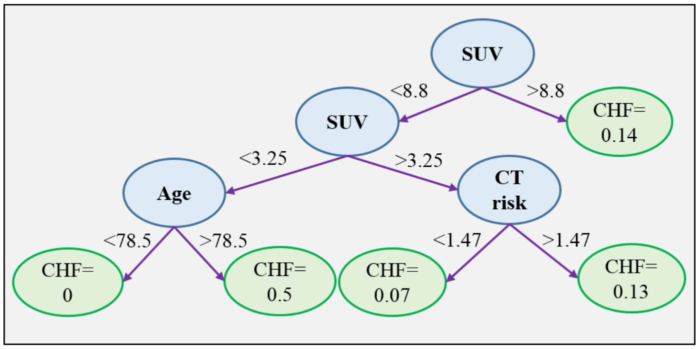


Figure 4.8: One of the trees obtained from the random survival forest (RSF) to predict the recurrence free survival (RFS). Following this tree, one is able to obtain the cumulative hazard function (CHF), for each patient, at a desired time point. Based on the CHFs, the survival can also be calculated.

that no recurrence event has been observed at this node. The RSF model, however, did not reveal any important predictor for the LC, and the concordance index could not be improved. Likewise, the RSF did not improve the accuracy of predicting the OS and DC, compared to the Cox PHM.

The 2-year risk score for RFS can be estimated by summing over the CHF values up to 2 years, obtained at discrete time points. This score can, specially, be used to compare the 2-year RFS risk scores of patients. For instance, for two patients, one of

Table 4.6: Results obtained from the proposed MDR-SURV and the single-scale methods.

Method	C-index
Proposed MDR-SURV framework	**73**%
First scale only (tumor region)	53%
Second scale only (10 pixels added to each side)	69%
Third scale only (20 pixels added to each side)	57%

which is censored after three years and 4 weeks, and the other has experienced the event of recurrence after one year and 3 weeks, the risk scores obtained from the RSF are 7.5 and 11.27, respectively.

Interpretability of the Deep Learning-based Features: To enhance interpretability of the extracted deep features (make them more tangible), we have conducted correlation analysis between the features extracted from the layer before the final layer in the DRTOP model and hand-crafted features, as shown in Figs. 4.9 and 4.10. In these heat maps, blue and red colors show positive and negative correlation, respectively. The darker the color, the stronger the relation. As it can be inferred from Figs. 4.9 and 4.10, features associated with the PET-risk are highly correlated with hand-crafted features extracted from PET images. The ones associated with the CT-risk, also, show correlation with some hand-crafted features extracted from CT images, although the correlation is not as strong as it is with the PET-risk.

Comparision with Previous Studies: Generally speaking, it is difficult to directly compare our study with previous works, as models are developed based on different datasets. Next, we focus on highlighting the differences between the proposed DRTOP architecture and previous relevant studies. Considerably lower than the obtained concordance index (c-index) of 68% using the proposed DRTOP model, the CNN model proposed by Zhu *et al.* [139] reaches a c-index of 62.9% in predicting OS in lung cancer patients, utilizing pathological images, which capture different information, compared to PET/CT images. Furthermore, the clinical parameters, such as SUV, and their predictive importance, are not considered in their study. The deep learning-based OS prediction model, developed by Wang *et al.* [160], reaching a c-index of 70%, also differs from DRTOP, in that multi-scale CT slices are utilized. Multi-scale input refers to including not only the tumor region itself, but also the surrounding tissues. Features extracted from the ROI (tumor) are not the only features that influence the outcome. Studies [134] have shown that the tissues surrounding the

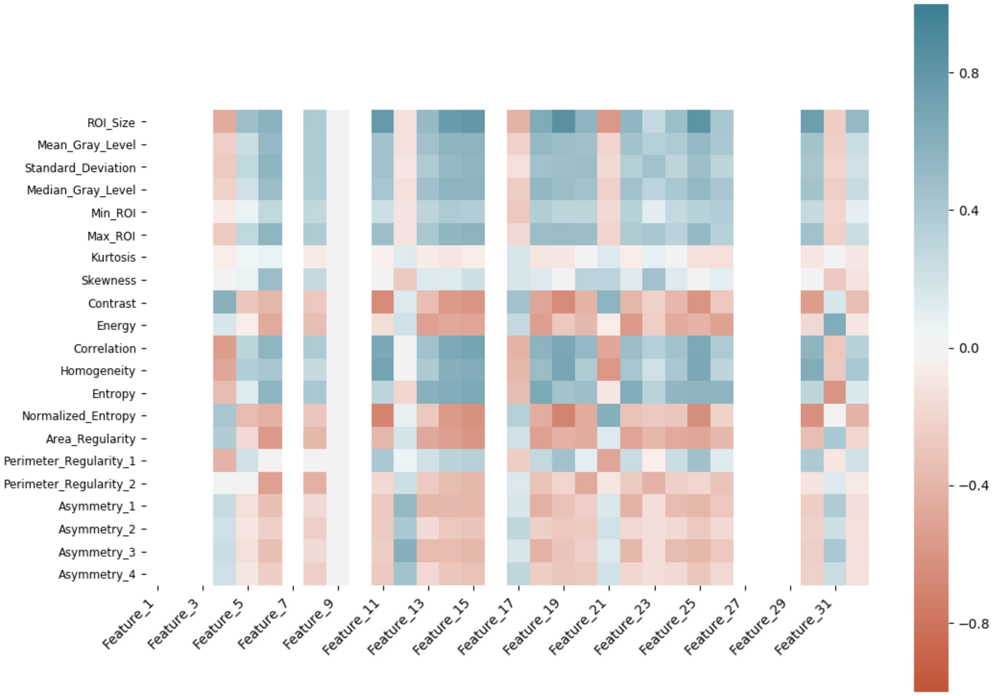


Figure 4.9: Correlation of the hand-crafted radiomics features extracted from PET images, with the deep learning-based features extracted from layer before the output layer of the model, trained on PET images.

tumor, also, play an important role in predicting the outcome. To be able to compare the predicted OS, using the DRTOP model, with the study by Wang *et al.*, we modified the DRTOP framework to account for multi-scale inputs, leading to a modified model referred to as MDR-SURV [9]. In paticular, we cropped the CT and PET slices from three different scales, shown in Fig. 4.11, where the first scale completely fits the tumor region, and the second and third scales are constructed by adding 10 and 20 pixels to each side of the tumor boundary, respectively. The three scales are stacked together, to form a 3-channel input, for both CT and PET scans. Other details of this modified architecture is quite similar to the DRTOP framework. The c-index, however, increases from 68% to 73%, which shows the significant importance of including multi-scale inputs. Our future plan is to study the impact of the surrounding regions of tumor on other time-to-event outcomes, including RFS, DC, and

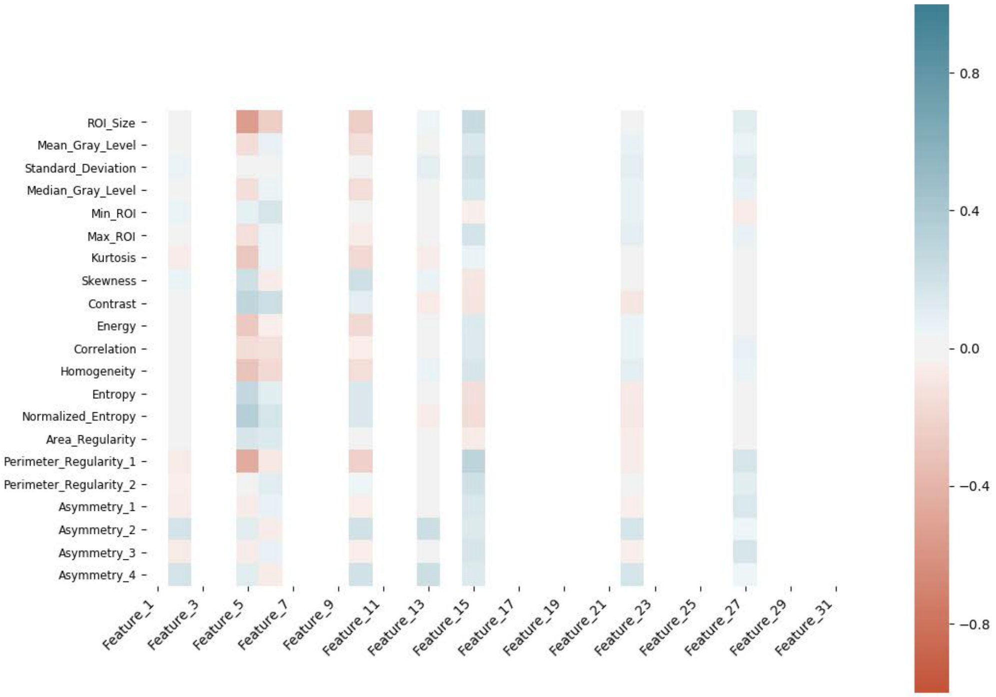


Figure 4.10: Correlation of the hand-crafted radiomics features extracted from CT images, with the deep learning-based features extracted from layer before the output layer of the model, trained on CT images.

LC. Table 4.6 shows the contribution of the different scales to the performance of the model. The proposed DRTOP framework leads to a better performance in predicting the OS, compared to the hand-crafted approach [19], increasing the c-index from 51% to 68%, because deep learning model is trained on its own, on the entirety of the image, as opposed to hand-crafted radiomics that are based on certain characteristics of the image.

PET risk and SUV are the only significant predictors of the RFS. However, there is much variability in the RFS that cannot be explained by the identified predictors, based on the Cox PHM. To the best of our knowledge, the prediction of lung cancer RFS, using deep learning, has not been, previously, investigated. The DC can be predicted by CT risk and SUV, leading to a concordance index of 63%. Deep learning-based DC prediction has been recently investigated by Xu *et al.* [141], where serial CT images are utilized to update the prediction, after each follow-up. This study, however, fails to provide high accuracy, having only the pre-treatment scans, without any follow-ups, which is the main goal of our work. The c-index using pre-treatment

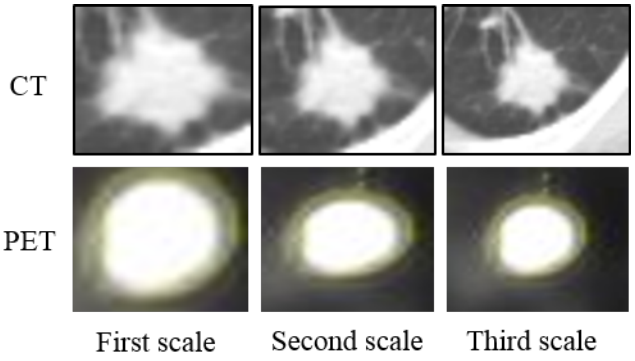


Figure 4.11: The CT and PET components cropped from three scales. These scales are stacked together and fed to the network.

images only, reported by Xu *et al.*, is 58.9% for 1 year distant control, and 58.5% for 2 year distant control. DRTOP and hand-crafted radiomics are almost on a par with each other, in predicting DC. While CT risk remains the only significant predictor of the LC, it does not lead to a satisfying concordance index. This means that there may be other factors influencing the LC. Hand-crafted radiomics completely failed to find predictors for the LC, as these features are extracted without considering the final goal, and there is no guarantee that they can contribute to the prediction.

It should be noted that all the results are reported based on the test set (20% of the whole dataset), and the low c-index obtained for LC, using the proposed DRTOP model, does not indicate a poor performance on the training set. In fact, our model was able to fit the training set and reach a high concordance index of 75%. Nevertheless, it failed to generalize well for the test set. This is the reason why the performance of the LC prediction was low. Increasing the number of patients may improve the performance.

As the Cox PHM is a semi-parametric model, thus, restricted to a predefined class of functions, we hypothesized that the poor performance may be due to an insufficiently met relationship between the predictors and the outcomes (RFS and LC). In other words, to ensure the appropriateness of the Cox PHM, the proportional hazards assumption must be met, which is not always the case. We, therefore, replaced the Cox PHM with a random survival forest (RSF) [154], which does not make this assumption, and calculated the importance values of the predictors, along with the final concordance indices. Our results demonstrated that, based on the RSF model, recurrence free survival (RFS) can be predicted with a concordance index of

64%, while all the predictors, except the radiation dose, showed high predictive value. Although a Cox PHM cannot predict the RFS, a non-parametric model can better explain the relation between the predictors and the outcome. Furthermore, although it is computationally expensive to calculate the cumulative hazard and absolute risk from the Cox PHM, as the baseline hazard is almost impossible to estimate, the RSF can be more easily used to provide the risk score. The RSF, however, may be biased, in the sense that it favors the variables with many split-points [161]. Variables with more split-points have a higher influence on the prediction error, and as such, they may be given more importance value.

4.4 Conclusion

In this chapter, we proposed a novel deep architecture, referred to as Deep learning-based radiomics for Time-to-event Outcome Prediction (DRTOP), consisting of two parallel CNNs, one of which was trained based on the CT component, and the other based on the PET component of the PET/CT. The output of the two models (referred to as CT and PET risks), together with clinical parameters such as Standardized Uptake Value (SUV), are fed to a Cox Proportional Hazards Model (PHM), to predict the time-to-event outcomes. The correlation between SUV and time-to-event outcomes has been previously studied, and it has been shown that SUV is of prognostic value for Overall Survival (OS), Local Control (LC), and Recurrence Free Survival (RFS). The SUV is, however, incapable of predicting the outcome, independently. To the best of our knowledge, this is the first time-to-event study that applies a deep learning method to the PET/CT images for staging lung cancer. Moreover, unlike most of the previous studies, which are limited to predicting the OS, our study explores the prediction of RFS, LC, and DC, which are of high clinical value. In conclusion, the proposed deep learning-based model on staging PET/CT images predicted the overall survival, recurrence free survival and distant metastasis in lung cancer patients. The comparison with hand-crafted radiomics showed that the deep learning model had a relatively better performance. While hand-crafted radiomics will continue to foster medical imaging research and give new insights about individual characteristics of medical images in patients with lung cancer, the combination of the two approaches may prove to be the future for clinical application. It should be noted that despite all the advancements in radiomics, there is still a long way until it is utilized as a stand-alone decision making tool. Challenges include the difficulty of acquiring rich amounts of training samples, considering the privacy issues and lack of homogeneous cohorts of patients, the difficulty of obtaining ground truth, unbalanced data, and image noise. The proposed model, however, can assist the radiologist with having a pick on factors and variables that are not available to the unaided human eye. In other words, deep learning-based radiomics may add complimentary predictive information in the personalized management of lung cancer patients.

Chapter 5

Deep Learning-based COVID-19 Diagnosis

As discussed in Chapter 1, COVID-19 diagnosis is the third application considered in this thesis. This chapter focuses on this problem considering different imaging modalities and clinical information. The chapter is organized as follows: Section 5.1 describes the intuition and problem statement, followed by the proposed COVID-CAPS framework in Section 5.2. Data collection and model development from CT scans are presented in Section 5.3. Finally, Section 5.4 concludes the chapter.

5.1 Introduction

Novel Coronavirus disease (COVID-19), first emerged in Wuhan, China [162], has abruptly and significantly changed the world as we know it at the end of the 2nd decade of the 21st century. COVID-19 seems to be extremely contagious and quickly spreading globally with common symptoms such as fever, cough, myalgia, or fatigue resulting in ever increasing number of human fatalities. Besides having a rapid human-to-human transition rate, COVID-19 is associated with high Intensive Care Unit (ICU) admissions resulting in an urgent quest for development of fast and accurate diagnosis solutions [162]. Identifying positive COVID-19 cases in early stages helps with isolating the patients as quickly as possible [163], hence breaking the chain of transition and flattening the epidemic curve.

Reverse Transcription Polymerase Chain Reaction (RT-PCR), which is currently the gold standard in COVID-19 diagnosis [162], involves detecting the viral RNA from sputum or nasopharyngeal swab. The RT-PCR test is, however, associated with relatively low sensitivity (true positive rate) and requires specific material and equipment, which are not easily accessible [162]. Moreover, this test is relatively time-consuming, which is not desirable as the positive COVID-19 cases should be identified and tracked as fast as possible [163]. Images [1] in COVID-19 patients, on the other hand, have shown specific findings, such as ground-glass opacities with rounded morphology and a peripheral lung distribution. Although imaging studies and theirs results can be obtained in a timely fashion, the previously described imaging finding may be seen in other viral or fungal infections or other entities such as organizing pneumonia, which limits the specificity of images and reduces the accuracy of a human-centered diagnosis.

Literature Review: Since revealing the potentials of computed tomography (CT) scans and X-ray images in detecting COVID-19 and weakness of the human-centered diagnosis, there have been several studies [164–166] trying to develop automatic COVID-19 classification systems, mainly using Convolutional Neural Networks (CNNs) [167]. Xu *et al.* [162] have first adopted a pre-trained 3D CNN to extract potential infected regions from the CT scans. These candidates are subsequently fed to a second CNN to classify them into three groups of COVID-19, Influenza-A-viral-pneumonia, and irrelevant-to-infection, with an overall accuracy of 86.7%. Wang *et al.* [163] have first extracted candidates using a threshold-based strategy. Consequently, for each case two or three regions are randomly selected to form the dataset. A pre-trained CNN is fine-tuned using the developed dataset. Finally, features are extracted from the CNN and fed to an ensemble of classifiers for the COVID-19 prediction, reaching an accuracy of 88%. CT scans are also utilized in Reference [168] to identify positive COVID-19 cases, where all slices are separately fed to the model and outputs are aggregated using a Max-pooling operation, reaching a sensitivity of 90%. In a study by Wang and Wong [169], a CNN model is first pre-trained on the ImageNet dataset [80], followed by fine-tuning using a dataset of X-ray images to classify subjects as normal, bacterial, non-COVID-19 viral, and COVID-19 viral infection, achieving an overall accuracy of 83.5%. In a similar study by Sethy and Behera [170], different CNN models are trained on X-ray images, followed by a Support Vector Machine (SVM) classifier to identify positive COVID-19 cases, reaching an accuracy of 95.38%.

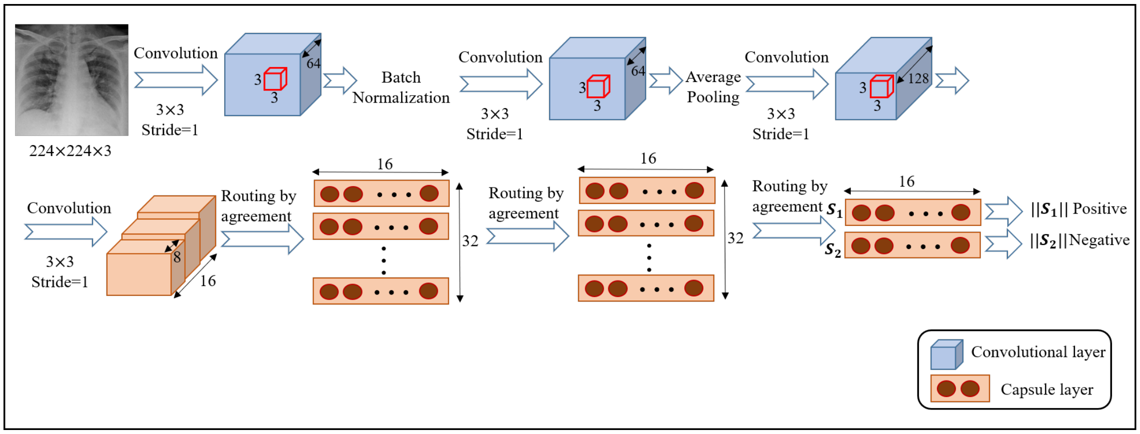


Figure 5.1: The proposed COVID-CAPS architecture.

Throughout this chapter, first COVID-19 diagnosis using X-ray images is considered. Since, CT scans contain more informative 3D features, we continue by developing deep learning models to identify COVID-19 from CT scans.

5.2 Capsule Networks for Identification of COVID-19 cases from Chest Radiographs (CXR)

In this section, we propose a Capsule Network-based framework, referred to as the COVID-CAPS [10], for COVID-19 identification using X-ray images. To potentially and further improve diagnosis capabilities of the COVID-CAPS, we considered pre-training and transfer learning using an external dataset of X-ray images, consisting of $94,323$ frontal view chest X-ray images for common thorax diseases. This dataset is extracted from the NIH Chest X-ray dataset [171] including $112,120$ X-ray images for 14 thorax abnormalities. It is worth mentioning that our pre-training strategy is in contrary to that of Reference [169] where pre-training is performed based on natural images (ImageNet dataset). Intuitively speaking, pre-training based on an X-ray dataset of similar nature is expected to result in better transfer learning in comparison to the case where natural images were used for this purpose.

The architecture of the proposed COVID-CAPS is shown in Fig. 5.1, which consists of 4 convolutional layers and 3 Capsule layers. The inputs to the network are 3D X-ray images. The first layer is a convolutional one, followed by batch-normalization. The second layer is also a convolutional one, followed by average pooling. Similarly, the third and forth layers are convolutional ones, where the forth layer is reshaped to form the first Capsule layer. Consequently, three Capsule layers are embedded in the COVID-CAPS to perform the routing by agreement process. The last Capsule layer contains the instantiation parameters of the two classes of positive and negative COVID-19. The length of these two Capsules represents the probability of each class being present.

Since we have developed a Capsule Network-based architecture, which does not need a large dataset, we did not perform any data augmentation. However, since the number of positive cases, N^+, are less than the negative ones, N^-, we modified the loss function to reduce the class imbalance effect. In other words, more weight is given to positive samples in the loss function, where weights are determined based on the proportion of the positive and negative cases, as follows

$$\text{loss} = \frac{N^+}{N^+ + N^-} \times \text{loss}^- + \frac{N^-}{N^+ + N^-} \times \text{loss}^+, \tag{5.1}$$

where loss^+ denotes the loss associated with positive samples, and loss^- denotes the loss associated with negative samples.

As stated previously, to potentially and further improve diagnosis capabilities of the COVID-CAPS, we considered pre-training the model in an initial step. In contrary to Reference [169] where ImageNet dataset [80] is used for pre-training, however, we constructed and utilized an X-ray dataset. The reason for not using ImageNet for pre-training is that the nature of images (natural images) in that dataset is totally different from COVID-19 X-ray dataset. It is expected that using a model pre-trained on X-ray images of similar nature would result in better boosting of the COVID-CAPS. For pre-training with an external dataset, the whole COVID-CAPS model is first trained on the external data, where the number of final Capsules is set to the number of output classes in the external set. From existing 15 disease in the external dataset, 5 classes were constructed with the help of a thoracic radiologist, with 18 years of experience in thoracic imaging (A. O.). To fine-tune the model using the COVID-19 dataset, the last Capsule layer is replaced with two Capsules to represent positive and negative COVID-19 cases. All the other Capsule layers are fine-tuned, whereas the conventional layers are fixed to the weights obtained in pre-training.

In summary, COVID-CAPS architecture contains the following modifications applied to the original Capsule Network presented in Reference [51]:

- The Capsule Network presented in Reference [51] originally works on a dataset of digital numbers, which are black-and-white and small in size compared to the X-ray images. To make the Capsule Network applicable in the problem at hand, we have extended the Capsule layers and the number of routing procedures to be able to extract useful patterns from X-ray images.

- The dataset originally used for the development of the Capsule Networks is completely balanced in terms of the number of instances available for each class label. The COVID-19 identification problem, however, is restricted to highly unbalanced datasets, as COVID-19 is a relatively new disease. To account for this unbalanced dataset, we modified the original margin loss to assign more penalty to mis-classified positive cases.

- We pre-trained the Capsule Network to compensate for the small available dataset. The pre-training is performed on an external dataset with 5 classes, reflected in 5 final Capsules. These 5 Capsules are then replaced with two, and all the Capsule layers are fine-tuned on the main COVID-19 dataset.

We used Adam optimizer with an initial learning rate of 10^{-3}, 100 epochs, and a batch size of 16. We have split the training dataset, into two sets of training (90%) and validation (10%), where training set is used to train the model and the validation set is used to select a model that has the best performance. Selected model is then tested on the testing set, for the final evaluation. The following four metrics are utilized to represent the performance: Accuracy; Sensitivity; Specificity, and Area Under the Curve (AUC).

COVID-CAPS Performance : To conduct our experiments, we used the same dataset as Reference [169]. This dataset is generated from two publicly available chest X-ray datasets [173, 180]. As shown in Fig. 5.2, the generated dataset contains four different labels, i.e., Normal; Bacterial; Non-COVID Viral, and; COVID-19. As the main goal of this study is to identify positive COVID-19 cases, we binarized the labels as either positive or negative. In other words, the three labels of normal, bacterial, and non-COVID viral together form the negative class.

Using the aforementioned dataset, the proposed COVID-CAPS achieved an accuracy of 95.7%, a sensitivity of 90%, specificity of 95.8%, and AUC of 0.97. The

Table 5.1: Results obtained from the proposed COVID-CAPS, along with the results from Reference [170]. Pre-trained CNN refers to a CNN with the same front-end as the COVID-CAPS

Method	Accuracy	Sensitivity	Specificity	Number of Trainable Parameters
COVID-CAPS without pre-training	95.7%	90%	95.8%	**295,488**
Pre-trained COVID-CAPS	**98.3%**	80%	**98.6%**	**295,488**
Reference [170]	95.38%	**97.29%**	93.47%	23,000,000
Pre-trained CNN	96.24%	50%	96.97%	368,508,226

Table 5.2: Description of the External X-ray images dataset used for pre-training COVID-CAPS.

Final Category	**Initial Categories**	**Number of Images**
No Findings	No Findings	60361
Tumors	Infiltration, Mass, Nodule	16103
Pleural Diseases	Effusion, Pleural Thickening, Pneumothorax	8042
Lung Infection	Consolidation, Pneumonia	1668
Others	Atelectasis, Cardiomegaly, Edema, Emphysema, Fibrosis, Hernia	8149

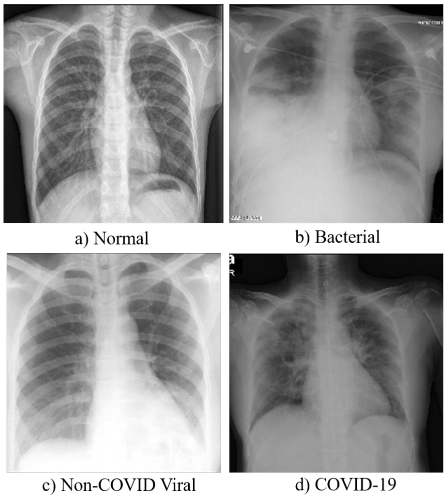

a) Normal
b) Bacterial
c) Non-COVID Viral
d) COVID-19

Figure 5.2: Labels available in the COVID-19 dataset.

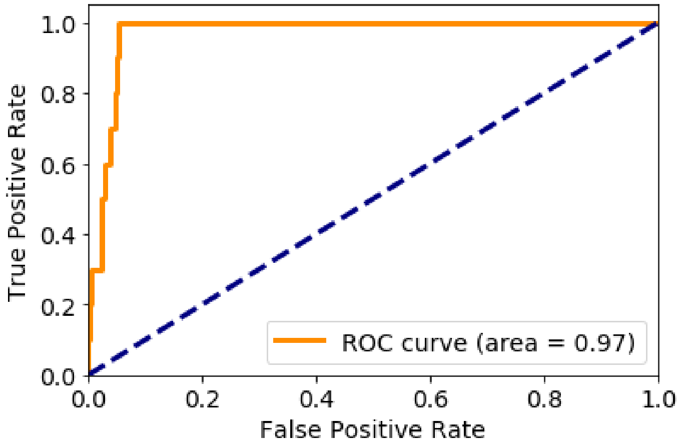


Figure 5.3: ROC curve from the proposed COVID-CAPS. Without pre-training refers to training from scratch.

obtained receiver operating characteristic (ROC) curve is shown in Fig. 5.3. *In particular, false positive cases have been further investigated to have an insight on what types are more subject to being mis-classified by COVID-19. It is observed that 54%*

of the false positives are normal cases, whereas bacterial and non-COVID cases form only 27% *and* 19% *of the false positives, respectively.*

As shown in Table 5.1, we compare our results with Reference [170] that has used the binarized version of the same dataset. COVID-CAPS outperforms its counterpart in terms of accuracy and specificity. Sensitivity is higher in the model proposed in Reference [170], that contains 23 million trainable parameters. Reference [165] is another study on the binarized version of the same X-ray images. However, as the negative label contains only normal cases (in contrast to including all normal, bacterial, and non-covid viral cases as negative), we did not compare the performance of the COVID-CAPS with this study. *It is worth mentioning that the proposed COVID-CAPS has only* $295,488$ *trainable parameters. Compared to* 23 *million trainable parameters of the model proposed in Reference [170], therefore, COVID-CAPS can be trained and used in a more timely fashion, and eliminates the need for availability of powerful computational resources.*

In another experiment, we pre-trained the proposed COVID-CAPS using an external dataset of X-ray images, consisting of $94,323$ frontal view chest X-ray images for common thorax diseases. This dataset is extracted from the NIH Chest X-ray dataset [171] including $112,120$ X-ray images for 14 thorax abnormalities. This dataset also contains normal cases without specific findings in their corresponding images. In order to reduce the number of categories, we classified these 15 groups into 5 categories based on the relations between the abnormalities in each disease. The first four groups are dedicated to No findings, Tumors, Pleural diseases, and Lung infections categories. The fifth group encompasses other images without specific relations with the first four groups. We then removed $17,797$ cases with multiple labels (appeared in more than one category) to reduce the complexity. The adopted dataset is then used to pre-train our model. Table 5.2 demonstrates our classification scheme and distribution of the data. Results obtained from fine-tuning the pre-trained COVID-CAPS is also shown in Table 5.1, according to which, pre-training improves accuracy and specificity. The ROC curve is shown in Fig. 5.3, according to which, the obtained AUC of 0.99 outperforms that of COVID-CAPS without pre-training.

Based on an inclusive study reported in Reference [174], human-centered COVID-19 detection from chest radiography leads to a high sensitivity, whereas specificity remains as low as 25%. The low specificity can lead to excessive expenses to isolate and treat false positive cases. The obtained specificity of 98.6% using the proposed COVID-CAPS can significantly assist radiologists to lower the number of reported false positives. Furthermore, the ROC curve can provide physicians with a means to calibrate and balance the sensitivity and specificity. In other words, by changing the probability threshold, above which the positive label is assigned to a subject, physicians are able to form the desired balance between sensitivity and specificity. To make this point more clear, we changed the probability threshold based on the ROC curve from the default value of 0.5 to 0.44. This new threshold increases the sensitivity to 100%, while specificity is slightly decreased to 98.4%.

To further elaborate on the effectiveness of the proposed model, we designed a CNN that has the same front-end as that of the COVID-CAPS. In other words, it has the same convolutional layers (the first four main layers of the COVID-CAPS). The Capsule layers, however, are replaced with three fully-connected layers, the first two of which have 256 neurons and the last one, having a Sigmoid activation, has two neurons representing the two classes of positive and negative COVID-19 cases. It is worth noting that we considered fully-connected layers after the front-end, because to some extent they resemble Capsule layers in the sense that there is no shared weights or kernels. This CNN is pre-trained on the same external dataset. In the fine-tuning phase, the convolutional layers are kept fixed and only the fully-connected layers are retrained. Furthermore, the cross-entropy loss function is modified (similar in nature to the modifications introduced on the margin loss of the COVID-CAPS in Eq. (5.1)) to give more penalty to mis-classified positive cases. All other hyper-parameters, including the optimizer and learning rate, exactly resemble the hyper-parameters of the COVID-CAPS. The training, validation, and test sets are also the same as the ones used in COVID-CAPS. Based on the obtained results, which are presented in Table 5.1, the designed CNN, having $368,508,226$ trainable parameters, achieves an accuracy of 96.24%, a sensitivity of 50%, and a specificity of 96.97%. The lower performance of the CNN, and the fact that it has exactly the same front-end with only the Capsule layers replaced with fully-connected ones support the effectiveness of the Capsule layer with the routing by agreement mechanism.

Finally, it is worth providing some intuition on COVID-CAPS time and space complexity. In particular, following the literature [175] we model the time complexity as a function of the number of required multiplications in both Capsule and fully-connected layers. Generally speaking, a fully-connected layer involves a matrix multiplication. Considering $m \times d_1$ and $n \times d_2$ neurons in two consecutive fully-connected layers, the required matrix multiplication involves $m \times d_1 \times n \times d_2$ multiplication operations. Reshaping the two fully-connected layers into two consecutive Capsule layers leads to m Capsules of dimension d_1 making predictions for n Capsules of dimension d_2. Each single prediction involves $d_1 \times d_2$ multiplications, as each lower layer Capsule i with dimension d_1 should be multiplied by the weight matrix $\boldsymbol{W}_{ij}$ to form the prediction $\hat{\boldsymbol{u}}_{j|i}$ for the higher layer Capsule j of dimension d_2. In other words, $\boldsymbol{W}_{ij}$ has d_1 rows and d_2 columns. Considering n Capsules in the lower layer and m Capsules in the higher layer, the total number of operations is $m \times d_1 \times n \times d_2$ which is exactly the same as the fully-connected scenario. However, each parent Capsule is calculated as a weighted average over the predictions. Weighting each prediction $\hat{\boldsymbol{u}}_{j|i}$ by the coupling coefficient c_{ij} involves d_2 (dimension of the prediction and parent Capsule) multiplications. Again having n Capsules in the lower layer and m Capsules in the higher layer, one routing by agreement process includes $d_2 \times n \times m$ multiplications. In conclusion, even with one round of routing by agreement, which means equal contribution of all the predictions, a Capsule layer has $d_2 \times n \times m$ multiplications more than a fully-connected layer. In practice, however, Capsule Networks require far less layers to have comparable performance with CNNs. To illustrate this point we calculated the time needed to predict the outcome of one single subject using the proposed COVID-CAPS. Our TITAN Xp GPU computer takes almost 0.16 seconds to calculate the outcome, whereas this time is approximately 1.62 seconds for the ResNet-50 model utilized in Reference [170]. Finally, regrading the space complexity, as we showed in the Table 5.1, COVID-CAPS contains far less trainable parameters compared to its counterparts. In particular, while trained COVID-CAPS occupies almost 1.5 Megabytes, the ResNet-50 requires 98 Megabytes.

5.3 Diagnosis of COVID-19 from CT scans

Although, CXR can act as a quantitative method to assess the extent of COVID-19 involvement and estimate the risk of Intensive Care Unit (ICU) admission, it still has lower sensitivity compared to Computed Tomography (CT) [176]. Due to high sensitivity and rapid access, chest CT plays a significant role in diagnosis and management of COVID-19 and has been recognized as the most sensitive imaging modality to detect complications [177]. It is worth noting that the developed imaging-based AI algorithms for the purpose of COVID-19 diagnosis can pave the path for the development of similar automatic systems for potential future pandemics, for which RT-PCR tests are not available.

5.3.1 COVID-CT-MD, COVID-19 computed tomography scan dataset applicable in machine learning and deep learning

Despite the high potential of CT in contributing to the COVID-19 research and clinical usage, publicly available datasets are mostly limited to a few number of cases, are not accompanied with other types of respiratory diseases to facilitate comparisons, and are not associated with suitable labels. Furthermore, cases may be collected from different sources with different imaging protocols, limiting a unified study. In a few identified datasets, available CT scans are limited to only infected slices, rather than the complete volume. Another important aspect that should be considered in the available datasets is that whether labels are available in a patient-level, slice-level, and lobe-level fashion. The later can further contribute to identify the location of the COVID-19 infection. Finally, different types of labels and information, suitable for different tasks, are provided in identified datasets. Table 5.3 provides an overview of the available datasets along with the provided COVID-19 related information.

Table 5.3: Available COVID-19 CT scan datasets. NA stands for not available.

	Number of cases			**Label type**		**Data Source**		**CT volume**		**Label Level**		
Dataset	COVID	CAP	Normal	Classification	Segmentation	Multiple	Single	Available	Not available	Patient-level	Slice-level	Lobe-level
Reference [178]	49	NA	NA		✓	✓		✓			✓	
Reference [179]	20	NA	NA		✓	✓		✓			✓	
Reference [180]	20	NA	NA		✓	✓		✓			✓	
Reference [181]	856	NA	254	✓		✓		✓		✓		
Reference [182]	216	NA	55	✓		✓			✓		✓	
Reference [183]	60	NA	60	✓		✓			✓		✓	
Reference [184]	95	NA	282	✓			✓	✓		✓	✓	
Reference [185]	2,980	NA	NA	✓		✓		✓		✓		
COVID-CT-MD	169	60	76	✓			✓	✓		✓	✓	✓

The introduced COVID-19 CT scan dataset, referred to as the COVID-CT-MD [11], is applicable in Machine Learning (ML) and deep learning studies of COVID-19 classification. In particular, COVID-CT-MD dataset consists of 169 confirmed positive COVID-19 cases (gathered from 2020/02/23 to 2020/04/21), 76 normal cases (gathered from 2019/01/21 to 2020/05/29), and 60 Community Acquired Pneumonia (CAP) cases (gathered from 2018/04/03 to 2019/11/24). All these cases are collected from Babak Imaging Center in Tehran, Iran, and labeled by three experienced radiologists in patient-level, slice-level, and lobe-level manners. Patient-level label refers to a single diagnosis assigned to the participant, whereas slice-level and lobe-level refer to identifying slices and lobes demonstrating infection, respectively. More importantly, the whole CT volume is available for all the participants. COVID-CT-MD is presented in Table 5.3, along with the previous datasets, to highlight its differences. Regarding Reference [184], we would like to mention that while this Reference provides only COVID-19 and normal cases, COVID-CT-MD provides CAP cases additionally. Furthermore, COVID-CT-MD is the only classification-related dataset that contains lobe-level information, which can significantly improve and contribute to the localization and analysis of the COVID-19 infection.

Data Collection Methods

This section provides a description of the data collection procedure, inclusion criteria, and de-identification. Furthermore, detailed statistics of the data is presented to facilitate its usage. More importantly, applicability of the COVID-CT-MD dataset for development of ML/DNN solutions is explained. This section is concluded by describing the possible limitations of the provided dataset.

Data Collection: The COVID-CT-MD dataset contains volumetric chest CT scans of 169 patients positive for COVID-19 infection, 60 patients with CAP, and 76 normal patients. COVID-19 cases are collected from February 2020 to April 2020, whereas CAP cases and normal cases are collected from April 2018 to December 2019 and January 2019 to May 2020, respectively, in Babak Imaging Center, Tehran, Iran. Three main criteria are considered by three radiologists for classifying the participants, as follows:

1. Imaging findings including:
    - Ground Glass Opacities (GGOs), referring to hazy transparent opacities;

    - Consolidation pattern, which means the air in the alveoli and peripheral bronchioles is replaced by fluid;
    - Crazy Paving, referring to thickened interlobular septa and intralobular lines superimposed on a background of ground-glass opacity;
    - Bilateral and multifocal lung involvement;
    - Peripheral distribution; and
    - More distribution in lower lobes.

2. Clinical findings including symptoms, characteristics, patient history, and RT-PCR outcome if available; and
3. Epidemiology, referring to whether the participant comes from high risk areas or has had close contact with a positive COVID-19 patient.

If a participant is identified positive according to all three criteria, COVID-19 label is assigned. Otherwise, the participant is classified as either CAP or normal. This procedure is followed by the three radiologists. Subsequently the majority voting is adopted for the final assignment. The three radiologists have 88.9% agreement in identifying COVID-19, CAP, and normal cases, whereas the first and second radiologists have 91.1% agreement, the first and third radiologists have 97.4% agreement, and the second and third radiologists have 89.1% agreement.

A subset of 54 COVID-19, and 25 CAP cases were analyzed by the first radiologist to identify and label slices with evidence of infection. The labeled subset of the data contains 4,957 number of slices demonstrating infection and 18,392 number of slices without infection.

Besides CT slices, clinical data is collected for the patients, which includes the following:

- Patients' age;
- Patients' gender;
- Patients' weight;
- Clinical characteristics: including symptoms, reason for scanning, and patients' history;
- Surgery history;

- Follow-up: some of the COVID-19 patients are followed-up after scanning and their status including recovery, hospital admission, and death is recorded;
- RT-PCR: positive RT-PCR outcome is available for some of the COVID-19 patients.

Table 5.4: CT scan settings used to acquire the COVID-CT-MD dataset.

Diagnosis	Slice Thickness (mm)	Peak Kilovoltage (kVp)	Exposure Time (ms)	X-ray Tube Current (mA)	SID (mm)	SOD (mm)	Exposure values (mAs)
COVID-19	2	110 – 130	600	153 – 343	940	535	61.2 – 180.0
CAP	2	110 – 120	420 – 600	94 – 500	940 – 1040	535 – 570	38.4 – 175.24
Normal	2	110	600	132 – 343	940	535	60.4 – 163.71

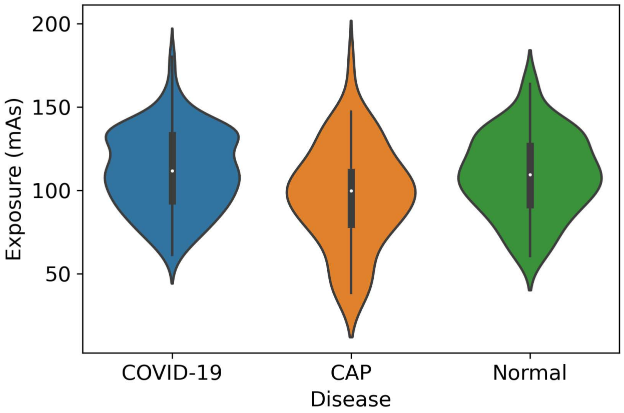


Figure 5.4: The distribution of the Exposure values for COVID-19, CAP and Normal cases.

Table 5.5: The statistical parameters (mean and standard deviation) of the Exposure values.

Diagnosis	Exposure mean	Exposure standard deviation
COVID-19	111.43	23.70
CAP	96.64	29.75
Normal	109.18	23.97

CT scans are comprised of cross-sectional 2D images from thin sections of the body (slices), creating a 3D representation of the structures inside the body. In the modern CT scanners, a rotating X-ray generator sends multiple X-ray beams into the object from multiple angles. The amount of the radiation passed through the object is then captured by sensitive radiation detectors, followed by a computer-assisted process, which reconstructs the information obtained from the detectors into detailed sequential images using image reconstruction techniques [186]. All images in COVID-CT-MD are obtained from a SIEMENS, SOMATOM Scope scanner in the axial view, using the helical acquisition technique, i.e., the patient is moved through the gantry while the X-ray beams and detectors are spinning rapidly around the patient. The images are reconstructed using the Filtered Back Projection (FBP) reconstruction method [187]. The reconstruction matrix size (output size of the images) is set to 512×512, and the D40s reconstruction kernel is used to reduce the blurring and noise by modifying the frequency contents of the data during the image reconstruction in the scanner [188]. Finally, all images are provided in the Hounsfield Unit and saved in the Digital Imaging and Communications in Medicine (DICOM) format. It is worth mentioning that following the recommended chest CT protocols for suspected cases or follow up of metastasis, bronchiectasis, interstitial lung disease and pulmonary infections [189] all images are Non-Contrast CT (NCCT) and none of them is CT Pulmonary Angiography (CTPA). Acquired images are, consequently, reconstructed into high resolution CT (HRCT).

Table 5.4 shows different CT acquisition settings, where Peak KiloVoltage (kVp) and Exposure Time affect the radiation exposure dose, while slice thickness represents the axial resolution. As shown in Table 5.4, slice thickness, kVP, and exposure time are almost the same with a few variations in a few CAP cases. Distance of Source to detector and Distance of Source to patient, which are traditionally referred to as SID and SOD, respectively, are also the same in all cases except for a few CAP cases. The minimum and maximum exposure value (in mAs) used in the scanning process is also presented in Table 5.4. The exposure value determines the total radiation dose in CT scan. The distribution of the exposure values is illustrated by the violin plots for each disease type in Figure 5.4. Accordingly, the mean and standard deviation of the exposure values are reported in Table 5.5.

CT Acquisition Care in The Medical Imaging Department: As COVID-19 is highly contagious, all the staff of the medical imaging department involved in the CT acquisition are provided with personnel protective equipment (PPE). More importantly, there is a minimum of 5-minute time slack between two consecutive CT scans, allowing enough time to sanitize the CT scanner.

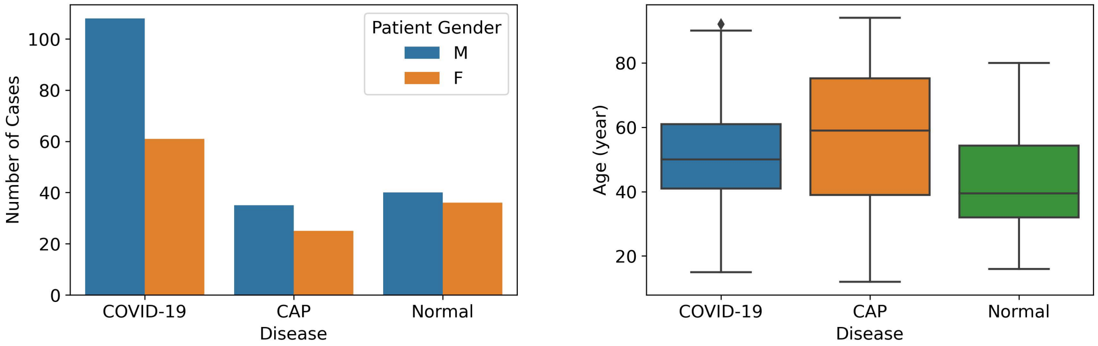
Data Inclusion and Exclusion Criteria: All cases with confirmed clinical diagnosis are included in the dataset. Nevertheless, during the data collection procedure, there were some cases related to the late 2019, with manifestations similar to those of COVID-19. However, as the first COVID-19 case in Iran is reported in early February 2020, these cases were excluded from the dataset. Furthermore, according to the radiologists' assessment, images with poor quality and visible artifacts were excluded. In summary, 320 cases were initially screened, among which 5% (15 cases) were excluded according to the radiologists' judgement, allowing 305 high quality CT studies.

De-identification: To respect the patients' privacy and comply with the DICOM supplement 142 (Clinical Trial De-identification Profiles) [190], we have de-identified all the CT studies by removing or obfuscating every names, UIDs, dates, times, comments, and center-related information. Some helpful DICOM attributes related to the patients' gender and age, the scanner type, and the image acquisition settings have been retained to preserve the statistical characteristics of the dataset. Patient's ID and UID attributes which are necessary to retain the consistency of the CT studies are replaced by new generated values which does not allow the identification of the patients.

Data Statistics: The demographic distribution of the dataset describing the gender and age distributions is illustrated in Table 5.6 and Figure 5.5. Please note that, no restrictions were imposed on the participants to indicate a binary response. As shown in Figure 5.5(a), males outnumbered females in this dataset. However, we would like to mention that although male cases are dominant, according to a recent study [191], there is no correlation between the CT score and participants'

Table 5.6: Gender and age distribution in COVID-CT-MD

Diagnosis	Cases	Gender	Age (year)
COVID-19	169	108 M/61 F	51.96 ± 14.39
CAP	60	35 M/25 F	57.7 ± 21.7
Normal	76	40 M/36 F	43.4 ± 14.1



(a)

(b)

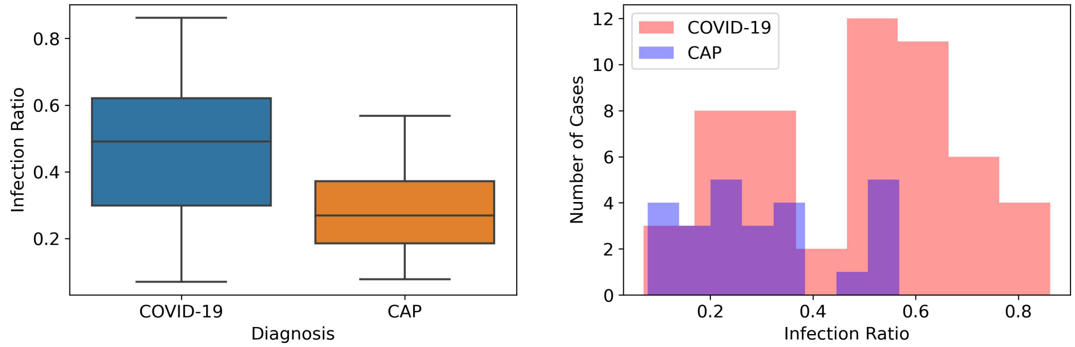
Figure 5.5: (a) The number of cases separated by the patient's gender. (b) The distribution of age for COVID-19, CAP and Normal cases.

Table 5.7: The number of cases, Slices, and Infection Ratio in the labeled dataset.

Diagnosis	Cases	Slices Demonstrating Infection	Slice without infection	Infection Ratio
COVID-19	54	3779	4269	7.0% – 86.2%
CAP	25	1178	2718	7.8% – 56.8%

gender. Furthermore, this dominance is common in most of the COVID-19-related datasets [177], possibly because men are more vulnerable to COVID-19, compared to women [192]. The boxplot in Figure 5.5(b) represents the important statistical parameters of the patients' age distribution. As shown in this boxplot, normal cases are mainly distributed in lower ages, while CAP cases are distributed in a wide range of ages with a higher average age. Regarding the ethnicity of the patients, the participants are Iranian (more than 60% Persian). Potential combination of the COVID-CT-MD dataset with other available ones, presented in Table 5.3, improves the applicability of AI algorithms to different populations.

As previously stated, part of the dataset is analyzed and the slice-level labels are extracted. The number of labeled cases and slices demonstrating infection are presented in Table 5.7. Infection ratio in this table represents the ratio of the slices demonstrating infection to the total number of slices in a CT scan, which varies for different cases based on the severity and stage of the disease. The minimum and maximum values for the infection ratio in the labeled dataset are presented in Table 5.7. The distribution of the Infection Ratio is also illustrated by the boxplots in



(a)

(b)

Figure 5.6: (a) The distribution of the Infection Ratio in the labeled dataset for COVID-19 and CAP cases. (b) The histogram of the Infection Ratio in the labeled dataset for COVID-19 and CAP cases.

Table 5.8: Number of cases and slices, respectively, demonstrating infection in each lobe. LLL: Left Lower Lobe – LUL: Left Upper Lobe – RLL: Right Lower Lobe and Lingula – RML: Right Middle Lobe – RUL: Right Upper Lobe

Diagnosis	LLL	LUL	RLL	RML	RUL
COVID-19	42&1669	38&1120	45&2008	26&420	29&826
CAP	13&374	5&117	18&519	7&186	9&208
Total	56&2079	43&1237	63&2527	33&606	38&1034

Figure 5.6(a), which demonstrate a higher infection ratio in COVID-19 cases compared to CAP cases. The histogram of the Infection Ratio values is illustrated in Figure 5.6(b).

In addition to the described slice-level labels, the detailed distribution of infection in each lobe of the lung is provided by the radiologists. Table 5.8 indicates the number of cases and slices with infection demonstrated in specific lung regions. Similar to Figure 5.6, where the infection ratio was presented for the total slices with infection in the lung, the average of lobe infection ratios are presented in Figure 5.7, illustrating the average ratio of slices demonstrating infection in a particular lobe to the total number of slices in a CT scan. As evident in Table 5.8 and Figure 5.7, the average infection ratio in the lower lobes is higher in both COVID-19 and CAP cases compared to other lung regions in our labeled dataset.

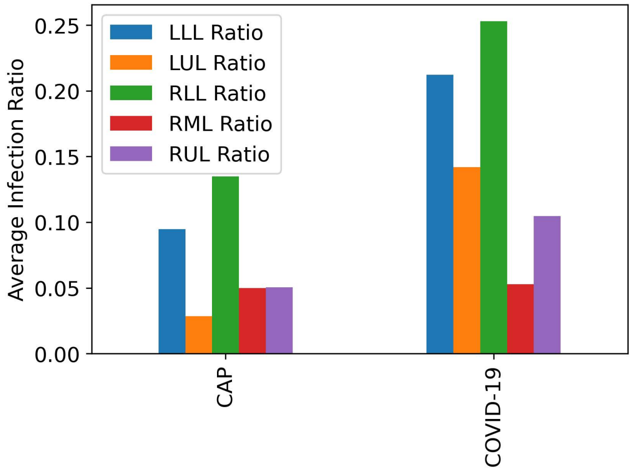


Figure 5.7: Average Infection Ratio in each lobe of the lung for COVID-19 and CAP cases in the labeled dataset.

Limitations: Although all cases and labels are confirmed by three experienced radiologists, we would like to describe a few limitations that the data users may encounter. These limitations are as follows:

- The slice and lobe labeling processes focus more on regions with distinctive manifestations rather than minimal findings.

- Not all the COVID-19 patients have confirmed positive RT-PCR result, as this test was not publicly accessible in Iran at the time of the first emergence of the COVID-19. Furthermore, the high load of patients in need of COVID-19 examination, did not allow for an inclusive RT-PCR test. The diagnosis of some patients in the COVID-CT-MD dataset is confirmed based on the CT findings, as well as the clinical results and epidemiology.

- Although most of the cases with low quality CT scans are excluded, there may still be some cases with mild motion artifact which is inevitable, since COVID-19 patients suffer from dyspnea.

- During the slice and lobe labeling process, some suspicious areas adjacent to the chest wall and diaphragm are not labeled as “infected”, due to their poor distinction.

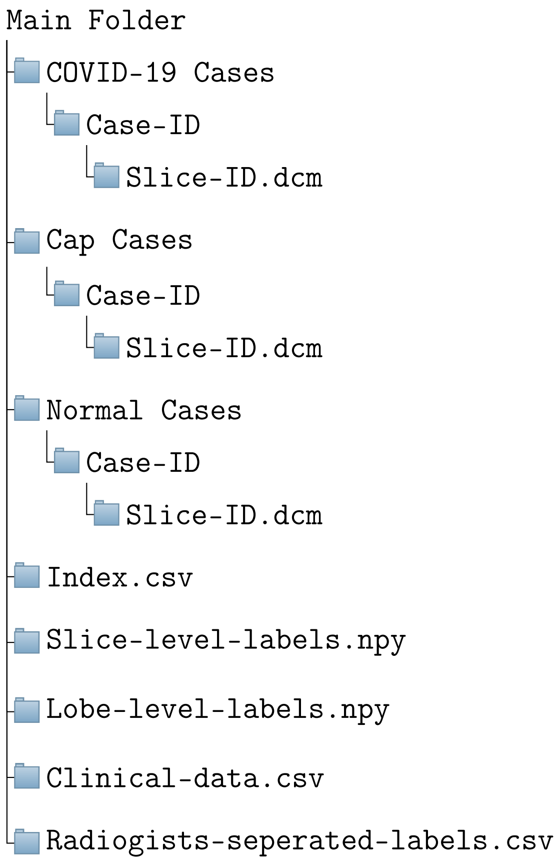


Figure 5.8: Structure of the data included in COVID-CT-MD dataset.

Data Records

The diagram in Figure 5.8 shows the structure of the COVID-CT-MD dataset. The COVID-CT-MD dataset is accessible through Figshare [193]. COVID-19, CAP and Normal participants are placed in separate folders, within which patients are arranged in folders, followed by CT scan slices in DICOM format. “Index.csv” is related to the patients having slice-level and lobe-level labels. The indices given to patients in “Index.csv” file are then used in “Slice-level-labels.npy” and “Lobe-level-labels.npy” to indicate the slice and lobe labels. “Slice-level-labels.npy” is a 2D binary Numpy array in which the existence of infection in a specific slice is indicated by 1 and the lack of infection is shown by 0. In “Slice-level-labels.npy”, the first dimension represents the case index and the second one represents the slice numbers. “Lobe-level-labels.npy” is a 3D binary Numpy array in which the existence of infection in a specific lobe and slice is determined by 1 in the corresponding element of the array. Like the slice-level array, in “Lobe-level-labels.npy”, the two first dimensions represent the case index and slice numbers respectively. The third dimension shows the lobe indices which are specified as follows:

- 0 : Left Lower Lobe (LLL)
- 1 : Left Upper Lobe (LUL)
- 2 : Right Lower Lobe (RLL)
- 3 : Right Middle Lobe (RML)
- 4 : Right Upper Lobe (RUL)

It is worth noting that CT slices are sorted based on the "Slice Location" value stored in the corresponding DICOM tag "(0020,1041) - DS - Slice Location". The slice-level and lobe-level labels are provided according to described slice order. The researchers, however, can re-arrange the slices using other CT attributes based on their preference, as long as they re-arrange the labels accordingly. The COVID-CT-MD dataset is also accompanied with the clinical data, stored in "Clinical-data.csv". Finally, to facilitate the inter-observer reliability studies, labels assigned by the three radiologists are separately provided in "Radiologists-separated-labels.csv".

Technical Validation: Two noteworthy parameters in the studies using CT scans are the quality control and calibration of the scanning device. The longest time period between the scanner auto-calibration and the study in the COVID-CT-MD dataset is 1 day, which ensures calibrated and accurate performance of the scanning device. Furthermore, there is an annual SIEMENS quality control that ensures the absence of ring artifacts in the acquired CT scans.

Usage Notes: With the increasing number of COVID-19 patients, healthcare workers are overwhelmed with a heavy workload, lowering their concentration for a proper diagnosis. Accurate and timely COVID-19 diagnosis, on the other hand, is a critical factor in preventing the disease transition, treatment, and resource allocation. Machine Learning (ML), in particular Deep Learning (DL) based on Deep Neural Networks (DNN), is shown to be practical and effective in COVID-19 diagnosis and severity assessment. The COVID-CT-MD dataset is specifically designed to facilitate application of ML/DL in COVID-19-related tasks. In particular, this dataset can be used towards:

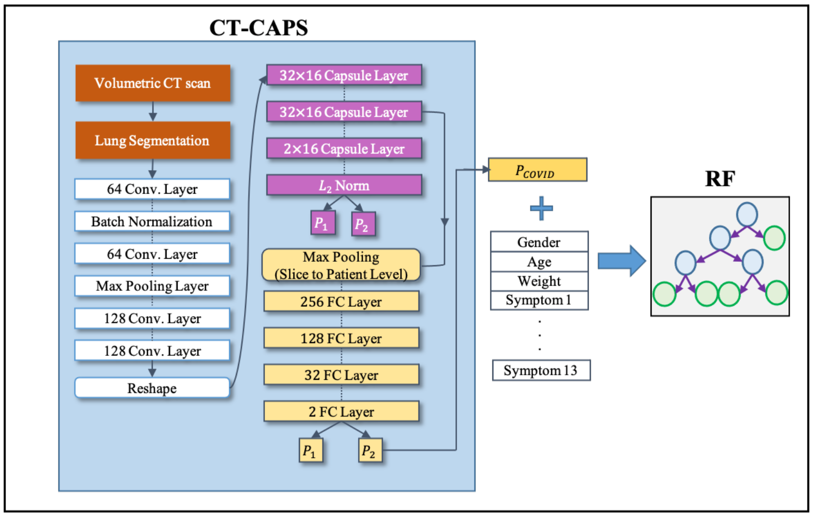


Figure 5.9: Pipeline of the proposed hybrid model for COVID-19 diagnosis.

- A patient-level binary classification [194] to distinguish COVID-19 from all other cases.

- A patient-level multi-class classification [194] to identify COVID-19, CAP, and normal participants.

- A slice-level [195] and lobe-level classification to separate infected slices and lobes from non-infected ones for further analysis.

- Slice-level and lobe-level labels can be used as additional inputs to segmentation models [196], to focus on only infected slices.

- Slice-level and lobe-level labels can be used in generative models to generate artificial COVID-19 images, towards increasing the security of the healthcare systems and developing attack resilient solutions [197].

5.3.2 Hybrid Deep Learning Model for Diagnosis of COVID-19 using CT Scans and Clinical/Demographic Data

In this section, we develop a hybrid deep learning model for COVID-19 diagnosis [12], using the COVID-CT-MD dataset, incorporating the patient's clinical/demographic data. The proposed hybrid model includes a CapsNet model as its building block, and a Random Forest (RF) Classifier. Fig. 5.9 depicts the pipeline and the main components of the proposed hybrid model.

CapsNet model: The CapsNet model is a fully automated framework, designed upon a stack of convolutional, pooling, batch normalization, and capsule network layers, outlined in Fig. 5.9, to extract slice-level feature maps from CT images in its first stage. The first stage is trained on the slice-level labeled portion of the data, with the goal of detecting COVID-19 suspected slices. The first stage generated feature maps go through a max pooling operation to develop one set of capsules per patient. The max pooling output forms the input to the second stage, which is a conventional multi-layer perceptron network with 256, 128, 32, and 2 layer sizes. The second stage is trained on the whole training dataset (patient-level label), and the final output is the probability of COVID-19 and non-COVID classes. The main advantage of the CapsNet framework over its counterparts is the far less number of trainable parameters ($0.5M$ compared to several millions of parameters used by other models) and its capability of capturing spatial relations between object in an image using the routing by agreement process. It also introduces a capsule network-based feature extraction framework to replace a large volumetric 3D CT scan by a very small matrix (32×16 in this case) which would be useful in many classification tasks working with volumetric image data.

It is worth noting how the inputs to the CapsNet are pre-processed. Since the chest CT images contain non COVID-related components and artifacts, we adopted a pre-trained U-Net-based model [198], in order to segment the lungs as the main regions of interest. This model is fine-tuned on a subset of COVID-19 images, to increase the robustness against infected regions. Furthermore, we have normalized the inputs between zero and one, to increase the generalizability. Following the literature [199, 200], we downsampled all slices from 512×512 to 256×256, which further reduces the time complexity, without loss of important details. At last, slices in which the lung region is not visible are removed.

Random Forest Classifier: An RF classifier consists of a set of T decision trees in which each tree $\mathcal{T}_t$, $t \in \{1, \ldots, T\}$, is trained separately on a randomly sampled subset of the training data. Each decision tree is comprised of several nodes, each of which represents a binary decision on a specific subset of the feature space. Each tree is comprised of an input (root) node, several internal (split) nodes and terminal (leaf) nodes. Each tree can be characterized independently based on the nodes by which it is constructed. Accordingly, Node j in a decision tree can be characterized by the split function parameter $\theta_j = (\phi_j, \psi_j, \tau_j)$, where ϕ_j denotes the feature selection function specifying which features from the full feature set are used in the split function associated with node j. Term ψ_j denotes the data separation function indicating which hyper-surface type is used to split the data, and τ_j is a threshold specifying the binary decision boundary at Node j. The axis-aligned hyper-plane strategy is used as the data separation function ψ_j, which splits the feature space of training samples using only one feature at a time by a hyper-plane which is associated with the feature axis. The training process aims to find the optimized split function parameter θ_j^* by maximizing the information gain (change in the entropy) objective function I given by

$$\theta_j^* = \underset{\theta_j}{\operatorname{argmin}} \, I_j, \tag{5.2}$$

where $I_j = I(S_j, S_j^L, S_j^R, \theta_j)$, S_j , S_j^L , and S_j^R represent training data before and after (Left/Right) split at node j respectively. Term I denotes the information gain function given by

$$I_j = H(S_j) - \sum_{i \in \{L,R\}} \frac{|S_j^i|}{|S_j|} H(S_j^i), \tag{5.3}$$

where H denotes the entropy function. The output of the RF classifier is calculated through an ensemble model combining output probabilities from all decision trees into a single output probability [201]. The average-based ensemble model is used in this study.

Hybrid Mechanism: The proposed hybrid model aims to integrate clinical and demographic data and the CT scans to improve the predictive performance and introduce another viewpoint to the model. As such, the patient-level COVID-19 probability scores P_{COVID}, acquired from the CapsNet framework trained on the CT scans, are concatenated with the encoded clinical and demographic data $\{gender, age, weight, s_1, \ldots, s_{13}\}$ and fed to the RF Classier. The patient's gender is encoded into 0 and 1 and $s_1, \ldots, s_{13}$ represent binary values corresponding to the patient's symptoms. The output of the RF classifier is the probability distribution associated with target classes of non-COVID (CAP and normal) and COVID-19, which is followed by a thresholding mechanism with the default value of 0.5 to provide the predicted classes.

Results from the Hybrid Model: In this study, we randomly sampled 60% of the dataset for training, 30% for testing, and 10% as the validation set to select the model with the minimum loss value during the training step. We made sure that the

Table 5.9: Patient-level classification results obtained from the proposed Hybrid Model and its counterparts.

Performance	**CapsNet**	**Proposed Hybrid Model**	**Proposed Hybrid Model (no lung)**	**Hybrid-CNN**	**Hybrid-Res50**
Accuracy(%)	89.8	**90.8**	80.6	78.6	81.6
Sensitivity(%)	**94.5**	**94.5**	89.1	87.3	**94.5**
Specificity(%)	83.7	**86.0**	69.8	67.4	65.1
AUC	**0.93**	0.92	0.89	0.85	0.90
# Params.	**0.5M**	**0.5M**	**0.5M**	$243.9M$	$24M$

Table 5.10: Features contributing the most to the final RF decision.

Rank	Feature	Importance Value
1	CapsNet output	6.73e-01
2	Age	1.15e-01
3	Weight	1.03e-01
4	Cough	1.89e-02
5	Fever	1.88e-02

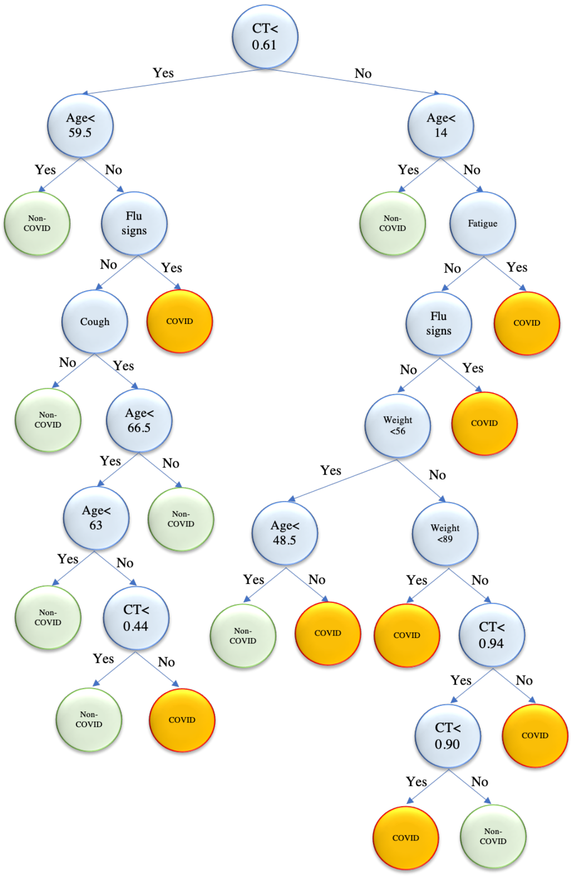


Figure 5.10: The internal structure of one of the decision trees created by the RF component of the proposed hybrid model. CT refers to the probability score obtained from the CapsNet model.

slices from the same patient are appeared in the same set to avoid any data leakage. The first stage of the CapsNet is trained using the Adam optimizer, with the learning rate of $1e-4$, batch size of 16, and 100 epochs. As the healthy slices outnumber the infected ones, we used balanced class weights. The second stage of the CapsNet is trained with the initial learning rate of $1e-3$, and 500 epochs. The RF classifier used for this study contains $1,000$ decision trees. The minimum number of samples required to split an internal node is 2, and the minimum number of samples required to be at a terminal (leaf) node is 1. The trees do not have a fixed depth and grow until all terminal nodes are pure (unable to divide the samples based on a set of features) or until all leaves contain less than 2 samples.

The performance of the proposed hybrid model on the described in-house dataset is presented in Table 5.9. As shown in Table 5.9, the proposed hybrid model is compared with several counterparts, the first of which is the CapsNet framework when the standard cutoff probability threshold of 0.5 is used. The second counterpart is the same hybrid model, with the difference that the whole CT image is used without lung segmentation, as an ablation study. Furthermore, we compared the proposed model with a Hybrid-CNN, which means the capsule layers are replaced with two fully connected layers with the size of 128. Finally, a Hybrid-Res50 is implemented for comparison where the Resnet50 [202] model is used as the feature extraction model in which the fully connected layer with 2,048 neurons before the last layer is taken as the feature map. As shown in Table 5.9, the proposed Hybrid model outperforms its counterparts, demonstrating the beneficial effects of the aggregation of clinical/demographic data and the CT scans, and incorporating a capsule network-based model.

The importance (i.e., contribution to the final decision) of the features is also extracted from the RF classifier and the 5 most important features with their corresponding importance value are listed in Table 5.10. In order to visualize the decision making procedure occurring inside the RF classifiers, the internal structure of one of the decision trees created by the proposed hybrid model is depicted in Fig. 5.10. The nodes and branches, which correspond to the split functions, features, and thresholds are demonstrated in this figure.

5.3.3 Human-level COVID-19 Diagnosis from Low-dose CT Scans Using a Two-stage Time-distributed Capsule Network

The main concern of widespread use of CT scan as a screening tool for suspected patients during the outbreak is the radiation exposure. In some scenarios, severely symptomatic patients will need multiple chest CT scans during the course of their disease. The cumulative effect of these multiple exposures can significantly increase the radiation dose. Studies [203] have shown that the projected radiation to body organs during chest CT scan is highest in thyroid, lung, breast, and esophagus. Due to their longer life expectancy, higher dose-effective breast tissue and cell proliferation [204, 205], children and young women are more vulnerable to radiation exposure damage with increased risk of radiation-following malignancy. As low as reasonably achievable (ALARA) [206] rule states that whenever radiation is expected, the exposure should be kept at the minimum achievable level such that the resulting scan still provides reasonable resolution.

Diagnostic accuracy of Low and Ultra-low-dose CT scan (LDCT and ULDCT) in detection and follow-up of pulmonary nodule and other lung pathologies has been previously established [207]. The radiation dose associated with standard chest CT is estimated at 7 mSv, which is reduced to 1-1.5 mSv with LDCT methods and as low as 0.3 mSv with ULDCT ones. The advantage of the low dose protocols is the reduction of radiation dose by more than 80%. Recent studies [208] have shown that DNA double-strand breaks and chromosome damage increased in patients undergoing a standard-dose CT scan while no effect on human DNA was demonstrated in patients undergoing low-dose CT scan. LDCT and ULDCT have shown significant accuracy in the detection of GGOs and consolidation in patients with pneumonia [209]. Since GGO and consolidation are the most common CT findings of COVID-19, recently, replacing standard CT scan with LDCT and ULDCT has been recommended [210] as a solution to decrease radiation exposure in COVID-19 patients. In a retrospective study [211], LDCT with iterative reconstruction (IR) demonstrated sensitivity, specificity, positive predictive value, negative predictive value, and accuracy of about 90% in the diagnosis of COVID-19. In conjunction with other clinical findings, LDCT and ULDCT can potentially replace standard-dose for the evaluation of patients, in particular pregnant and young women, and pediatric populations, to decrease radiation exposure [212].

To the best of our knowledge, Reference [213] is the only study considering LDCT in AI-based COVID-19 analysis, by simulating standard dose scans from low dose ones. The aforementioned study, however, uses synthesized data, i.e., it does not use real LDCT/ULDCT data from COVID-19 individuals, and does not deal with the disease diagnosis, which is the main focus of our research. We hypothesize that AI can achieve a human-level performance in diagnosing COVID-19 based on LDCT and ULDCT scans.

We developed a two-stage deep learning model, shown in Fig. 5.11, built upon the capsule network architecture, which takes segmented lung regions as inputs. The first stage will provide a subset of candidate slices to be analyzed in the next stage, which focuses only on the disease type. To train the first stage, we used 2D CT images and their corresponding label (infectious vs non-infectious) to construct a slice-level classifier whose output determines the probability of the input image belonging to a specific target class (infectious vs non-infectious). We then extracted 10 [214] slices with the highest infection probability for each patient to be used as the input of the second stage. The architecture of the first stage initiates with a stack of four convolutional layers, one pooling layer, and one batch normalization layer which are augmented by two shortcut connections to deliver shallow features to the deeper layers of the model. These layers are then followed by a stack of three capsule layers to generate the final output, which is the probability of the input image belonging to the related target class. It is worth noting that in the first stage, we are dealing with an imbalanced dataset with more number of slices without the evidence of infection. To cope with this imbalanced dataset, we have modified the loss function in the training step and considered a higher penalty for the errors in the slices demonstrating infection. The second stage of the proposed AI framework is a time-distributed capsule network that takes the 10 candidates from the previous stage as inputs. These images are processed in parallel through capsule networks with the same architecture sharing all the trainable weights. These capsule networks consist of three convolutional layers, one batch normalization and one max pooling layer. The output of the last convolutional layer is reshaped to form the primary capsules, which then go through two capsule layers. The final capsule layer for each candidate corresponds to the three classes of COVID-19, CAP, and normal. To take into account the probability of the candidate slice being infected, COVID-19 and CAP classes are multiplied by the infectious probability generated by the first stage. The normal class is also multiplied by one minus the infectious probability. At the end, a global max pooling operation is applied to the outputs of the capsule networks corresponding to candidate slices, to make the final decision. We trained the second stage time-distributed capsule network with an Adam optimizer with learning rate of $1e^{-4}$, batch size of 8, and 150 epochs. Similar

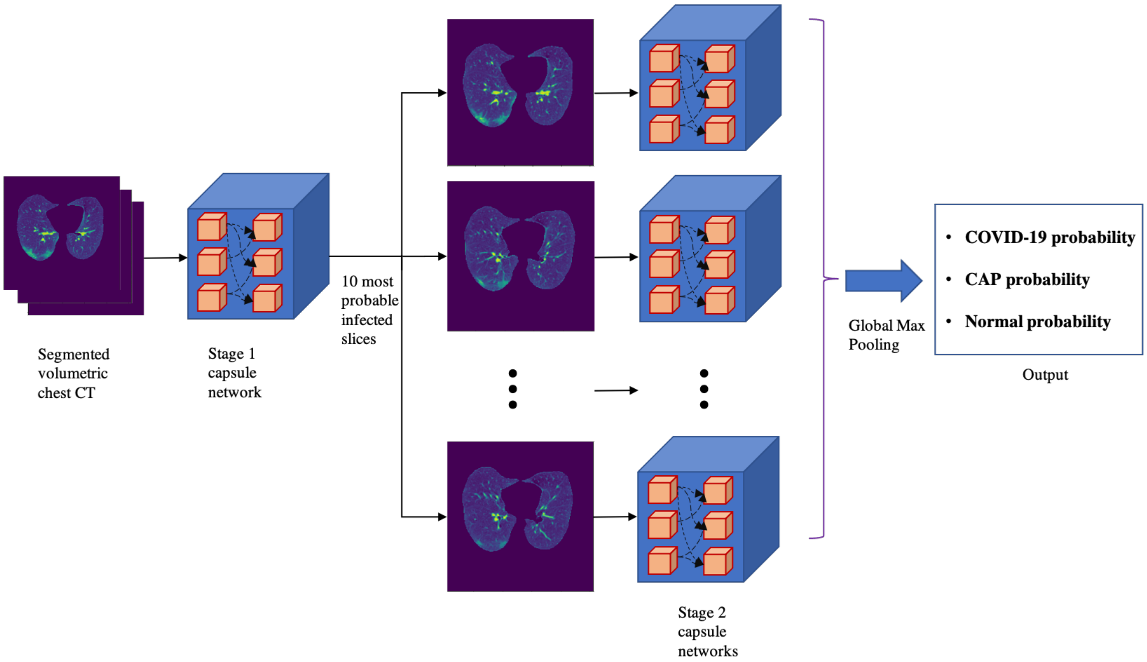


Figure 5.11: The proposed 2 stage deep learning model for COVID-19 diagnosis using LDCT/ULDCT. At the first stage, CT slices go through a capsule network, one by one, to detect those with evidence of infection. At the second stage, 10 most probable slices with infection detected in the previous stage go through a time-distributed capsule network, output of which determines the probability of COVID-19, CAP, and normal, after applying a global max pooling.

to the first stage, we used a modified margin loss function to consider more penalty for the minority class.

After training the two-stage deep learning model, output probabilities of the three classes (COVID-19, CAP, normal) are concatenated with the 8 clinical data (demographic and symptoms, i.e., sex, age, weight, and presence or absence of 5 symptoms of cough, fever, dyspnea, chest pain, and fatigue) and fed to a multi-layer perceptron (MLP) model, shown in Fig. 5.12. This model has 4 fully-connected layers with 64 neurons, where each layer is followed by batch normalization. The last layer includes 3 neurons with a “Softmax" activation function. We trained the MLP model with a cross-entropy loss and Adam optimizer with the learning rate of $1e^{-4}$, batch size of 16, and 500 epochs.

We collected an in-house dataset of LDCT and ULDCT scans of 104 COVID-19 positive cases, and 56 normal cases, collected in October 2020, December 2020, and January 2021, Babak Imaging Center, Tehran, Iran. Diagnosis of 36.5% of the

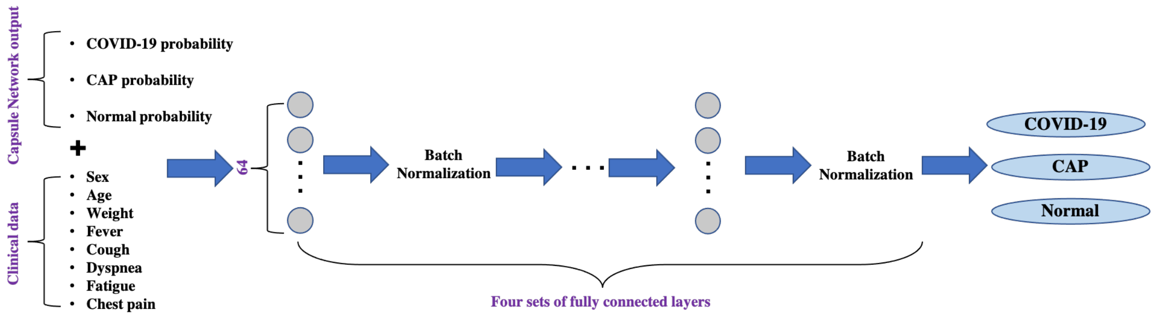


Figure 5.12: The MLP model combining the output of the two-stage deep learning model with the clinical data. Clinical data includes demographic characteristics and 5 common COVID-19 and CAP symptoms. Four sets of fully connected layers determine the final output.

Table 5.11: Characteristics of the in-house dataset. SD stands for standard deviation.

	COVID-19	CAP	Normal	P-value: COVID-19 vs. rest	P-value: CAP vs. rest	P-value: Normal vs. rest
Sex: Men	58.6%	58%	39.3%	0.7386	0.1848	0.0314
Age in Years (Mean ± SD)	49.53 ± 15.5	57.78 ± 21.94	40.18 ± 15.37	0.7283	0.0003	0.0002
Weight in Kg (Mean ± SD)	80.75 ± 14.84	67.38 ± 12.96	75.91 ± 14.52	0.0001	0.0000	0.6881
Dyspnea	26.9%	18%	45%	0.8932	0.0091	0.0425
Cough	31.7%	53%	33.93%	0.1480	0.0160	0.4916
Fever	14.4%	36%	9%	0.5130	0.0242	0.0589
Chest Pain	7%	0%	10.7%	0.7571	0.9999	0.6439
Fatigue	10.5%	0%	1.7%	0.0107	0.9999	0.2681

COVID-19 cases (38 cases) is confirmed with the RT-PCR test. The rest are specified by taking the consensus between 3 experienced thoracic radiologists, who labeled the dataset by taking the imaging findings, clinical characteristics (symptoms and history), and epidemiology into account. The three radiologists reached an agreement of 95.6%. They also scored the severity of the COVID-19 cases between 1 and 4, based on the percentage of the lung involvement. Four positive COVID-19 cases do not reveal any related imaging findings. As we did not have access to LDCT scans of CAP patients, we combined this dataset with 60 standard-dose volumetric CT scans [11]. The dataset characteristics are shown in Table 5.11. P-values are obtained using logistic regression, by considering three binary scenarios of COVID-19 versus CAP and normal, CAP versus COVID-19 and normal, and normal versus COVID-19 and CAP. Finally, a fourth experienced thoracic radiologist, blind to the ground-truth, labeled the collected dataset to compare the performance of the AI model with a human expert. The radiologist was first provided with only the CT scans, and then the clinical data.

To decrease bias towards a specific test set, we adopted a 10-fold cross validation approach [215] to assess the performance of the radiologist and the AI model, based on two scenarios of using CT scans only, and incorporating the clinical data. The dataset is randomly split into 10 equal size test sets, leading to 10 sets each including 22 cases. We made sure that each set contained 10% of the COVID-19, CAP, and normal cases, leading to 10 or 11 COVID-19, 6 CAP, and 5 or 6 normal cases in each test set. The AI model is trained 10 times, setting one of the test sets aside and using the rest for training. Averaging over the 10 folds, the slice-level classifier in the first stage achieved accuracy of 89.88%, sensitivity of 88.24%, and specificity of 92.01%, in detecting the slices with infection.

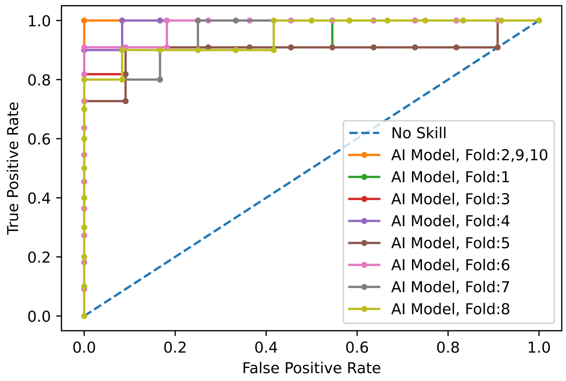


Figure 5.13: ROC curve for COVID-19 diagnosis (vs CAP and normal) using the proposed deep learning model and CT scans only.

Using only CT scans, we evaluated the developed deep learning model and compared it with the fourth thoracic radiologist, as shown in Table 5.12. Averaging over all the 10 folds, AI model achieves COVID-19 sensitivity of 89.5% ± 0.11, CAP sensitivity of 95% ± 0.11, normal sensitivity (specificity) of 85.7% ± 0.16, and accuracy of 90% ± 0.06. The radiologist, on the other hand, achieves COVID-19 sensitivity of 89.4% ± 0.12, CAP sensitivity of 88.33% ± 0.11, normal sensitivity (specificity) of 100%, and accuracy of 91.8% ± 0.07. We tested the hypothesis of the AI model and radiologist having the same performance, in term of accuracy, using a McNemar [216] test with the significance level of 0.05, leading to P-values over the significance level for all the 10 folds. The lower specificity of the AI model conforms the non-specific COVID-19 findings [217]. COVID-19 sensitivity versus one minus specificity is plotted in the receiver operating characteristics (ROC) curve, shown in Fig. 5.13. Area under the curve (AUC) is 0.96 ± 0.03.

Based on the CT scans only, we analyzed the misclassified COVID-19 cases through

Table 5.12: Performance of the AI model and the radiologist blind to the labels, using only CT scans.

	COVID-19 Sensitivity		**CAP Sensitivity**		**Normal Sensitivity**		**Accuracy**			
Fold	AI	Radiologist	AI	Radiologist	AI	Radiologist	AI	Radiologist	Accuracy P-value	COVID-19 AUC
1	$\frac{10}{11}$	$\frac{10}{11}$	$\frac{6}{6}$	$\frac{5}{6}$	$\frac{4}{5}$	$\frac{5}{5}$	$\frac{20}{22}$	$\frac{20}{22}$	1	0.95
2	$\frac{10}{10}$	$\frac{10}{10}$	$\frac{6}{6}$	$\frac{6}{6}$	$\frac{6}{6}$	$\frac{6}{6}$	$\frac{22}{22}$	$\frac{22}{22}$	1	1
3	$\frac{10}{11}$	$\frac{9}{11}$	$\frac{6}{6}$	$\frac{4}{6}$	$\frac{3}{5}$	$\frac{5}{5}$	$\frac{19}{22}$	$\frac{18}{22}$	1	0.91
4	$\frac{10}{10}$	$\frac{10}{10}$	$\frac{4}{6}$	$\frac{6}{6}$	$\frac{5}{6}$	$\frac{6}{6}$	$\frac{19}{22}$	$\frac{22}{22}$	0.25	0.99
5	$\frac{8}{11}$	$\frac{10}{11}$	$\frac{5}{6}$	$\frac{5}{6}$	$\frac{5}{5}$	$\frac{5}{5}$	$\frac{18}{22}$	$\frac{20}{22}$	0.5	0.9
6	$\frac{10}{11}$	$\frac{10}{11}$	$\frac{6}{6}$	$\frac{5}{6}$	$\frac{5}{5}$	$\frac{5}{5}$	$\frac{21}{22}$	$\frac{20}{22}$	1	0.98
7	$\frac{8}{10}$	$\frac{10}{10}$	$\frac{6}{6}$	$\frac{5}{6}$	$\frac{5}{6}$	$\frac{6}{6}$	$\frac{19}{22}$	$\frac{21}{22}$	0.625	0.96
8	$\frac{7}{10}$	$\frac{8}{10}$	$\frac{6}{6}$	$\frac{5}{6}$	$\frac{5}{6}$	$\frac{6}{6}$	$\frac{18}{22}$	$\frac{19}{22}$	1	0.95
9	$\frac{10}{10}$	$\frac{6}{10}$	$\frac{6}{6}$	$\frac{6}{6}$	$\frac{5}{6}$	$\frac{6}{6}$	$\frac{21}{22}$	$\frac{18}{22}$	0.375	1
10	$\frac{10}{10}$	$\frac{10}{10}$	$\frac{6}{6}$	$\frac{6}{6}$	$\frac{5}{6}$	$\frac{6}{6}$	$\frac{21}{22}$	$\frac{22}{22}$	1	1
Total	89.5% ± 0.11	89.4% ± 0.12	95% ± 0.11	88.33% ± 0.11	85.7% ± 0.16	100%	90% ± 0.06	91.8% ± 0.07	-	0.96 ± 0.03

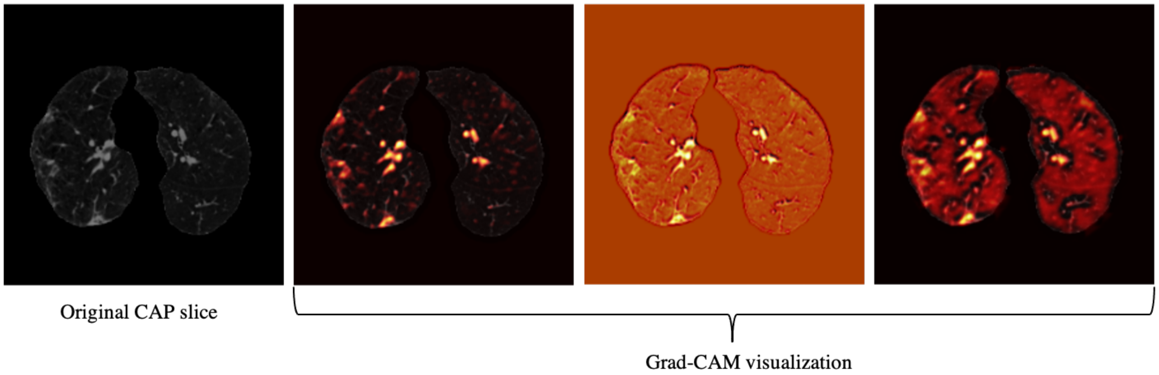


Figure 5.14: Grad-CAM visualization of one CAP slice. This figure shows that the proposed AI model is paying attention to relevant locations of the image.

all folds (11 cases in total), and studied their relation with the disease severity, coming to the conclusion that 4 out of 11 cases, did not have any related imaging findings, 5 were scored 1 by the three radiologists, one was scored 2, and only one case was scored at 3, which means the developed model is less likely to misclassify severe cases. Neither the developed model nor the experienced radiologist was able to detect the 4 COVID-19 cases without imaging findings, using CT scans only. Furthermore, since the CAP patients come from a different cohort and scanned with a standard dose, we visualized the model's output for CAP cases, one of which is shown in Fig. 5.14, using Grad-CAM localization technique [218]. This figure shows that the model is paying more attention to disease-related regions of the image, rather than dose-related ones. We performed the same localization technique on two slices with infection of the same COVID-19 patient, shown in Fig. 5.15.

Using both CT scans and clinical data, we evaluated the developed deep learning model and compared it with the radiologist, as shown in Table 5.13. Averaging over all the 10 folds, AI model achieves COVID-19 sensitivity of 94.3% ± 0.05, CAP sensitivity of 96.7%±0.07, normal sensitivity (specificity) of 91%±0.09 , and accuracy of 94.1%±0.03. The radiologist, on the other hand, achieves COVID-19 sensitivity of 94.4%±0.05, CAP sensitivity of 93.3%±0.08, normal sensitivity (specificity) of 100%, and accuracy of 95.4%±0.03. We tested the hypothesis of the AI model and radiologist having the same performance, using LDCT and clinical data, in terms of accuracy, leading to P-values over the significance level for all the 10 folds. COVID-19 sensitivity versus one minus specificity is plotted in the receiver operating characteristics (ROC) curve, shown in Fig. 5.16. Area under the curve (AUC) is 0.96 ± 0.03.

Table 5.13: Performance of the AI model and the radiologist blind to the labels, using both CT scans and clinical data.

	COVID-19 Sensitivity		**CAP Sensitivity**		**Normal Sensitivity**		**Accuracy**			
Fold	AI	Radiologist	AI	Radiologist	AI	Radiologist	AI	Radiologist	Accuracy P-value	COVID-19 AUC
1	$\frac{11}{11}$	$\frac{10}{11}$	$\frac{5}{6}$	$\frac{6}{6}$	$\frac{5}{5}$	$\frac{5}{5}$	$\frac{21}{22}$	$\frac{21}{22}$	1	0.99
2	$\frac{10}{10}$	$\frac{10}{10}$	$\frac{6}{6}$	$\frac{6}{6}$	$\frac{6}{6}$	$\frac{6}{6}$	$\frac{22}{22}$	$\frac{22}{22}$	1	1
3	$\frac{10}{11}$	$\frac{10}{11}$	$\frac{6}{6}$	$\frac{6}{6}$	$\frac{4}{5}$	$\frac{5}{5}$	$\frac{20}{22}$	$\frac{21}{22}$	1	0.91
4	$\frac{9}{10}$	$\frac{10}{10}$	$\frac{5}{6}$	$\frac{6}{6}$	$\frac{6}{6}$	$\frac{6}{6}$	$\frac{20}{22}$	$\frac{22}{22}$	0.5	0.98
5	$\frac{10}{11}$	$\frac{10}{11}$	$\frac{6}{6}$	$\frac{5}{6}$	$\frac{5}{5}$	$\frac{5}{5}$	$\frac{21}{22}$	$\frac{20}{22}$	1	0.97
6	$\frac{10}{11}$	$\frac{10}{11}$	$\frac{6}{6}$	$\frac{5}{6}$	$\frac{4}{5}$	$\frac{5}{5}$	$\frac{20}{22}$	$\frac{20}{22}$	1	0.99
7	$\frac{9}{10}$	$\frac{10}{10}$	$\frac{6}{6}$	$\frac{5}{6}$	$\frac{5}{6}$	$\frac{6}{6}$	$\frac{20}{22}$	$\frac{21}{22}$	1	1
8	$\frac{9}{10}$	$\frac{10}{10}$	$\frac{6}{6}$	$\frac{5}{6}$	$\frac{6}{6}$	$\frac{6}{6}$	$\frac{21}{22}$	$\frac{21}{22}$	1	0.98
9	$\frac{10}{10}$	$\frac{8}{10}$	$\frac{6}{6}$	$\frac{6}{6}$	$\frac{5}{6}$	$\frac{6}{6}$	$\frac{21}{22}$	$\frac{20}{22}$	1	0.95
10	$\frac{10}{10}$	$\frac{10}{10}$	$\frac{6}{6}$	$\frac{6}{6}$	$\frac{5}{6}$	$\frac{6}{6}$	$\frac{21}{22}$	$\frac{22}{22}$	1	0.99
Total	94.3% ± 0.05	94.4% ± 0.05	96.7% ± 0.07	93.3% ± 0.08	91% ± 0.09	100%	94.1% ± 0.03	95.4% ± 0.03	-	0.96 ± 0.03

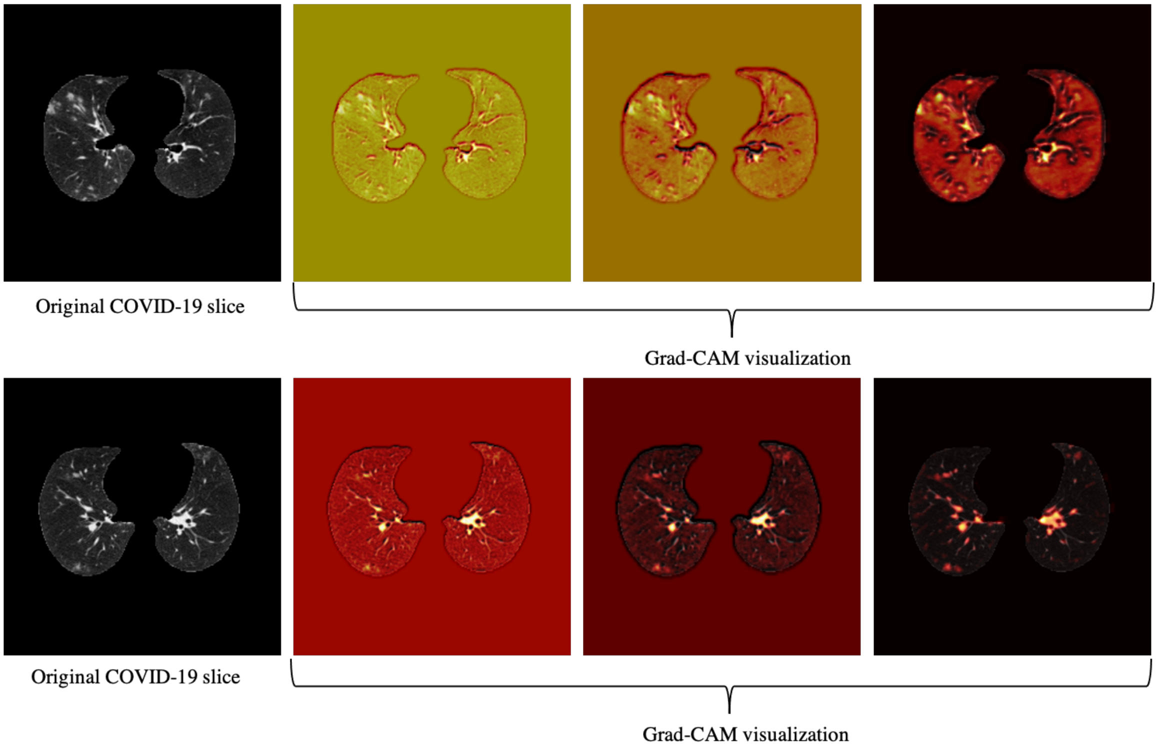


Figure 5.15: Grad-CAM visualization of two COVID-19 slices. This figure shows that the proposed AI model is paying attention to relevant locations of the image.

Based on using both CT scans and clinical data, we analyzed the misclassified COVID-19 cases through all folds (6 cases in total), and studied their relation with the disease severity, coming to the conclusion that 3 out of 6 cases, did not have any related imaging findings, one was scored 1 by the three radiologists, and two cases were scored at 3. Incorporating the clinical data, the AI model can detect one of the four positive COVID-19 cases, without having related imaging findings, whereas the radiologist did not detect any of them.

Finally, we tested the developed AI model, incorporating LDCTs and clinical data, on an extra set of 100 positive COVID-19 patients, whose diagnosis are confirmed with RT-PCR test and are collected in a different time interval (narrow validation). These patients were not included in any of the 10 folds and are completely unseen to the model and radiologist. While 68 out of 100 cases have imaging findings, 32 do not reveal any related manifestations. Male cases constitute 53% of the total cases, and age average is 46.16 with a standard deviation of 14.07. The AI model correctly identifies all the 68 positive cases having imaging findings, whereas it detects only 3 of

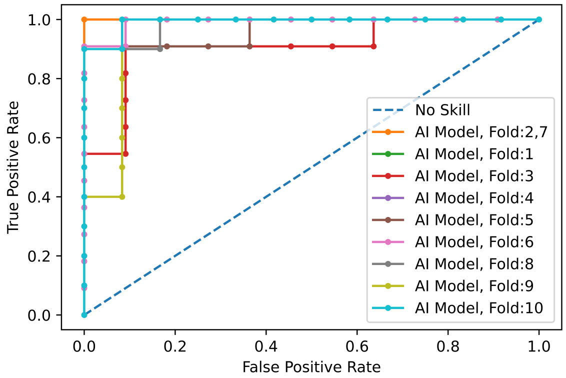


Figure 5.16: ROC curve for COVID-19 diagnosis (vs CAP and normal) using the proposed deep learning model and both CT scans and clinical data.

those not having related findings. Radiologist, on the other hand, correctly classifies 64 out of 68 patients having imaging findings as COVID-19 and classifies 4 as CAP. None of the cases without imaging findings are identified by the radiologist. The p-value between the AI model and radiologist's sensitivity is 0.01.

Although LDCT and ULDCT can reveal COVID-19 related findings and reduce the potential radiation-related harms, an accurate diagnosis requires full investigation by radiologists, which may not be possible during the outbreak. Based on our experiments, the proposed capsule network-based AI model has the potential to rapidly distinguish COVID-19 cases from CAP and normal ones with a human-level performance using LDCT and ULDCT, having a radiation dose of a single X-ray image. In other words, with minimal radiation, the developed AI system can assist the radiologists and contribute to controlling the chain of COVID-19 transmission. Furthermore, we showed that by incorporating the clinical data, COVID-19 sensitivity increases by 4.8%, CAP sensitivity increases by 1.7%, and normal sensitivity and accuracy increase by 5.3% and 4.1%, respectively.

Our study has some limitations. First, the dataset is collected from a single centre, and experiments are required to verify its performance on data from external institutes, as it is critical to investigate if the model generalizes to diverse population [219, 220]. Vulnerability to data shifts, and bias against underrepresented population [219] are also crucial to address before the AI model can be put into practice. It is worth mentioning that as the extra set of 100 positive COVID-19 patients are collected in a disjoint time interval from the original set, it can act as a narrow validation [220]. It is, however, collected from the same institute and thus does not account for broad validation. It is also of high interest to explore domain validation for COVID-19 diagnosis, where test set comes from different variants. Second, the sample size is relatively small. Verifying the model's performance on larger multi-centre datasets is the goal of our upcoming studies. The capsule network used in our study, is capable of handling small datasets compared to conventional models and due to fewer trainable parameters it is less prone to over-fitting, however, larger datasets can still improve the performance of the model. We also aim at expanding the proposed AI model to predict the disease severity besides the diagnosis. Moreover, although as shown in Figs. 5.14 and 5.15 visualization of the AI model's output shows it is paying attention to relevant regions, more research is required to increase its explainability. Low performance on COVID-19 cases without imaging finding is another limitation of the developed model.

5.4 Conclusion

The consequent global COVID-19 crisis has directed many research studies towards developing rapid and automated frameworks aiming to prevent the spread of the disease and flatten the epidemic curve. In this chapter, we proposed a Capsule Network-based framework, referred to as the COVID-CAPS, for diagnosis of COVID-19 from X-ray images. The proposed framework consists of several Capsule and convolutional layers, and the lost function is modified to reduce for the class-imbalance effect. The obtained results show that the COVID-CAPS has a satisfying performance with a low number of trainable parameters. Pre-training was able to further improve the accuracy, specificity, and AUC. Furthermore, capitalizing on the advantages of CT scans over X-ray images, we collected a rich dataset of CT scans from COVID-19, CAP, and Normal cases, along with their clinical characteristics. This dataset, referred to as COVID-CT-MD, is further utilized to develop a hybrid model incorporating patients' clinical/demographic data into an automated COVID-19 identification framework. We showed that adding such informative data to the models that are working only with the CT scans will improve the classification performance and increase the explainability of the obtained results. We utilized a Random Forest classifier to propose decision making strategies based on the large set of input features and identify the most informative ones. The proposed hybrid model achieved the accuracy of 90.8%, sensitivity of 94.5%, and specificity of 86.0% showing improvement over the original CapsNet framework which only relies on the features extracted from CT scans. Furthermore, we showed that the probability scores generated by the CapsNet framework along with the age, weight, cough symptom, and fever have the most influence on the confirmation of the COVID-19 infection. In addition, we demonstrated that even self-reported symptoms are beneficial and relatively reliable in the diagnosis of COVID-19 and CAP infections.

Although CT scans have shown considerable image findings related to COVID-19 diagnosis, they are associated with harmful impacts on the body. Low dose CT scans, on the other hand, can contribute to diagnosis with less harmful effects. In this sense, we collected a dataset of LDCT and ULDCT, based on which we proposed a time-distributed deep learning model for COVID-19 diagnosis. The developed AI model achieves human-level performance by incorporating LDCT/ULDCT and clinical data, having the advantage of reducing the risks related to radiation exposure. This model can act as a decision support system for radiologists and help with controlling the transmission chain. As our developed AI model is not intended to be a primary diagnostic tool, we aim at testing the model alongside a thoracic radiologist to assess its performance as a decision support tool rather than a stand-alone system.

Chapter 6

Conclusion and Future Direction

During the past decades, medical imaging made significant advancements leading to the emergence of automatic techniques to extract information that are hidden to human eye. Nowadays, the extraction of quantitative or semi-quantitative features from medical images, referred to as radiomics, can provide assistance in clinical care especially for disease diagnosis/prognosis. Throughout this Ph.D. thesis, we first presented an integrated sketch on radiomics by introducing its practical applications and processing modules. Then we focused on three important applications of radiomics, i.e., tumor classification, time-to-event-outcome prediction, and COVID-19 diagnosis. In the following, we first summarize the thesis contributions, presented in Table 6.1, and then discuss potential directions for future research.

6.1 Summary of Contributions

The thesis contributions can be summarized as follows:

- **Tumor Classification:** In Chapter 3, we discussed the most important application of the radiomics, which is the tumor type classification, and investigated the use of newly proposed Capsule networks for the problem of brain tumor classification. Since these networks can handle small number of training samples, and units in these networks are equivalent, they outperform CNNs in tumor classification problem. We followed-up this study, by presenting a CapsNet architecture that incorporates both the raw MRI brain images and the tumor coarse boundaries in order to classify the tumors. The proposed CapsNet architecture has two main advantageous: (i) First, the need for tumor exact annotation is eliminated, and; (ii) Second, it helps the CapsNet to focus on the main area, and at the same time, consider its relation with surrounding tissues. Our results show that the proposed approach is capable of increasing the classification accuracy, compared to the previous CapsNets and CNN architectures. We further improved the CapsNet ability to classify tumors, by proposing a 3D multi-scale network, capable of distinguishing between benign and malignant nodules. Capitalizing on the potentials of ensemble machine learning techniques, we proposed a boosting as well as mixture of CapsNets, and showed how the CapsNet itself can be viewed as a mixture of experts framework. Furthermore, to capture the model uncertainty, we developed a Bayesian workflow, capable of outputting both mean predictions and an index of uncertainty, thus referring the uncertain predictions to the human expert.

- **Time-to-event Outcome Prediction:** After the tumor classification problem, we focused on the time-to-event outcome prediction in Chapter 4, which is of high importance for treatment design. We presented a deep learning-based architecture that takes the multi-scale PET and CT images as inputs, and calculates the PET and CT risks. The CT risk, PET risk, age, gender, SUV, and radiations dose are then fed to a COX PHM, as well as an RSF, to predict the desired outcome. Our experiments on an in-house dataset of 132 patients show that the proposed framework outperforms its counterparts.

- **COVID-19 Diagnosis:** Last but not least, in Chapter 5, we focused on the emerging problem of COVID-19 diagnosis, by first developing a deep learning model to predict COVID-19 from X-ray images. Capitalizing on the advantages of CT scans over X-ray images, we collected a rich dataset of CT scans, along with clinical and demographic features, using which we developed a hybrid deep learning model to diagnose COVID-19. To reduce the radiation exposure associated with standard-dose CT scan, we collected an in-house dataset of LDCT and ULDCT and showed that a time-distributed deep learning model achieves human-level performance in COVID-19 diagnosis, based on LDCT, ULDCT, and clinical data.

Table 6.1: Summary of the contributions.

	Target	**Description**	**Advantages**
BoxCaps [2, 3]	Brain tumor classification	A capsule network fed with MRI images and rough tumor boundaries.	Segmentation is not required and handles small datasets.
BoostCaps [4]	Brain tumor classification	A boosted capsule network to improve performance.	No need to explore the space of possible architectures.
BayesCap [5]	Brain tumor classification	A Bayesian capsule network to handle uncertainty.	Return uncertain predictions to human expert.
3D-MCN [4]	Lung nodule classification	A 3D multi-scale Capsule network.	Captures features from the surrounding tissues and is fed with 3D input.
MIXCAPS [7]	Lung nodule classification	A mixture of capsule networks as individual experts.	Improves performance by splitting samples between experts.
DRTOP [8, 9]	Time-to-event outcome prediction	A CNN to predict lung cancer patients' survival and tumor recurrence.	Predicts response to treatment to avoid unnecessary procedures.
COVID-CAPS [10]	COVID-19 diagnosis	A Capsule network fed with X-ray to diagnose COVID-19	Handles small and unbalance dataset with less computation.
Hybrid COVID-19 model [12]	COVID-19 diagnosis	A Hybrid deep learning model to diagnose COVID-19 from CT scan and clinical information.	Combines information from two sources to have an inclusive decision.
LDCT [13]	COVID-19 diagnosis	A time-distributed capsule network to diagnose COVID-19 from LDCT.	Decreases radiation dose.

6.2 Future Direction

Our future directions are as follows:

- **Knowledge Distillation:** Practical and clinical application of the proposed deep learning models are limited by the fact that they require extensive computational resources. One solution is to adopt a knowledge distillation framework, in which the developed models are distilled into student models. These models are less complicated, networks which are aimed at producing the same output as the teacher models. Leveraging scalable models, such as EfficientNet [221], is another technique to develop more practical frameworks.

- **Aleatoric Uncertainty:** Including the aleatoric uncertainty is another important part of our future direction, in which the uncertainty of the data and its inherited noise is considered, besides the model's uncertainty. This type of uncertainty is of paramount importance in medical imaging domains, where clean data is rare and difficult to acquire.

- **Fusion Methodologies:** Properties of medical images such as their contrast and resolution varies significantly from one institute to another (from one dataset to another), because each institute may use different types of scanners and/or use different technical parameters. Development of novel and innovative information fusion methodologies together with construction of unifying schemes are critical to compensate for lack of standardization in this field and produce a common theme for comparing the radiomics results. Furthermore ground truth and annotations provided by different experts can vary significantly as experts, depending on their area of specialty (such as oncology and surgery), may consider and look for different details and landmarks in an image.

- **Image Noise:** Dealing with image noise is another challenging problem, which is common in all multi-media domains, but it is more severe in radiomics as there may be more unpredictable sources of variation in medical imaging. As an example, patient's breathing in the CT scanner can cause change of lung tumor location in consecutive slices bringing about difficulty in extracting stable radiomics features. Therefore, to achieve reliable personalized diagnosis and treatment, careful strategies should be developed to address the effects of these kinds of variations. Furthermore, there are several factors, such as imaging environments, capabilities of the scanners and other shortcomings of radiological images (e.g., radiations during acquisition or noisy acquisitions), that limit the resolution of the obtained medical images. For instance, the range of the captured frequencies is limited by the maximum sampling rate of the scanner, and increasing the rate, will increase the resolution, at the cost of an increased noise. Since access to high-quality images is necessary to achieve an early and accurate diagnosis/detection, there is an ongoing research on improving the quality of the medical images via development of advanced computational models to overcome the aforementioned shortcomings. One of such computational techniques is known as "Super-Resolution", aiming at reconstructing a high-resolution image, using several low-resolution instances. Deep learning networks, and CNNs in particular, are widely used in super-resolution problems, and so far have shown promising results.

- **Multi-centre Studies:** Most of the datasets we used are collected from single centres, and experiments are required to verify performance on data from external institutes, as it is critical to investigate if the model generalizes to diverse population. Vulnerability to data shifts, and bias against underrepresented population are also crucial to address before the AI models can be put into practice. It is also of high interest to explore domain validation for COVID-19 diagnosis, where test set comes from different variants.

- **Performance Lower Bound**: Generally speaking, deep learning models are verified based on one or more validation sets, and there is no guarantee on the performance lower bound. Posterior Cramer–Rao Bounds (CRB) [222] is one potential technique to model such guarantee.

- **Histopathological Images:** Lastly, our future direction contains an extensive research on histopathological images for disease diagnosis. These types of images contain detailed molecular information. They are however, very large in size (millions of pixels) and current models cannot handle them at once. Processing these images require specific multi-instance learning frameworks to split images without loss of information.

Bibliography

[1] P. Afshar, A. Mohammadi, K. N. Plataniotis, A. Oikonomou, H. Benali, “From Handcrafted to Deep Learning-based Cancer Radiomics: Challenges and Opportunities”, *IEEE Signal Processing Magazine*, vol. 36, no. 4, pp. 132-160, 2019.

[2] P. Afshar, A. Mohammadi, K. N. Plataniotis, “Brain tumor type classification via capsule networks”, *IEEE International Conference on Image Processing (ICIP)*, pp. 3129-3133, 2018.

[3] P. Afshar, K. N Plataniotis, A. Mohammadi, “Capsule Networks for Brain Tumor Classification Based on MRI Images and Coarse Tumor Boundaries,” *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1368-1372, 2019.

[4] P. Afshar, K. N Plataniotis, A. Mohammadi, “BoostCaps: A Boosted Capsule Network for Brain Tumor Classification,” *42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 11075-1079, 2020.

[5] P. Afshar, K. N Plataniotis, A. Mohammadi, “BayesCap: A Bayesian Approach to Brain Tumor Classification Using Capsule Networks,” *IEEE Signal Processing Letters*, vol. 27, pp. 12024-2028, 2020.

[6] P. Afshar, A. Oikonomou, F. Naderkhani, P.N. Tyrrell, K. Farahani, K. N Plataniotis, A. Mohammadi, “3D-MCN: A 3D Multi-scale Capsule Network for Lung Nodule Malignancy Prediction,” *Scientific Reports*, vol. 10, 2020.

[7] P. Afshar, F. Naderkhani, A. Oikonomou, M. J. Rafiee, A. Mohammadi, K. N. Plataniotis, “MIXCAPS: A Capsule Network-based Mixture of Experts for Lung Nodule Malignancy Prediction,” *Pattern Recognition*, vol.116, 2021.

[8] P. Afshar, A. Mohammadi, P.N. Tyrrell, P. Cheung, A. Sigiuk, K. N Plataniotis, E. Nguyen, A. Oikonomou, “DRTOP: Deep learning-based Radiomics for the Time-to-event Outcome Prediction in lung cancer,” *Scientific Reports*, vol. 10, 2020.

[9] P. Afshar, A. Oikonomou, K. N Plataniotis, A. Mohammadi, “MDR-SURV: a Multi-scale Deep learning-based Radiomics for SURVival prediction in pulmonary malignancies,” *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2013-2017, 2020.

[10] P. Afshar, Sh. Heidarian, F. Naderkhani, A. Oikonomou, K. N. Plataniotis, A. Mohammadi, “COVID-CAPS: A capsule network-based framework for identification of COVID-19 cases from X-ray images,” *Pattern Recognition Letters*, vol. 138, pp. 638-643, 2020.

[11] P. Afshar, Sh. Heidarian, , F. Naderkhani, M. J. Rafiee, A. Oikonomou, K. Samimi, K. N. Plataniotis, A. Mohammadi, “COVID-CT-MD: COVID-19 Computed Tomography (CT) Scan Dataset Applicable in Machine Learning and Deep Learning,” *Scientific Data*, vol. 8, 2021.

[12] P. Afshar, Sh. Heidarian, F. Naderkhani, M. J. Rafiee, A. Oikonomou, K. N. Plataniotis, A. Mohammadi, “Hybrid Deep Learning Model for Diagnosis of COVID-19 using CT Scans and Clinical/Demographic Data,” Accepted in *IEEE International Conference on Image Processing (ICIP)*, 2021.

[13] P. Afshar, M. J. Rafiee, F. Naderkhani, Sh. Heidarian, N. Enshaei, A. Oikonomou, F. Babaki Fard, R. Anconina, K. N. Plataniotis, K. Farahani, A. Mohammadi, “Human-level COVID-19 Diagnosis from Low-dose CT Scans Using a Two-stage Time-distributed Capsule Network,” *arXiv:2105.14656v1*, 2021.

[14] P. Afshar, A. Shahroudnejad, A. Mohammadi, K. N. Plataniotis, “CARISI: Convolutional Autoencoder-Based Inter-Slice Interpolation of Brain Tumor Volumetric Images,” *2018 25th IEEE International Conference on Image Processing (ICIP)*, pp. 1458-1462, 2018.

[15] P. Afshar, K. N Plataniotis, A. Mohammadi, “Capsule Networks’ Interpretability for Brain Tumor Classification Via Radiomics Analyses,” *2019 IEEE International Conference on Image Processing (ICIP)*, pp. 13816-3820, 2019.

[16] P. Lambin, E. Rios-velazquez, *et al.*, “Radiomics: Extracting more Information from Medical Images using Advanced Feature Analysis,” *European Journal of Cancer*, vol. 48, no. 4, pp. 441-446, 2012.

[17] V. Kumar, Y. Gu, *et al.*, “Radiomics: The Process and the Challenges,” *Magnetic Resonance Imaging*, vol. 30, no. 9, pp. 1234-1248, 2012.

[18] R. Gillies, P. Kinahan, *et al.*, “Radiomics: Images Are More than Pictures, They Are Data.," *Radiology*, vol. 278, No. 2, pp. 563-577, 2016.

[19] A. Oikonomou, F. Khalvati, *et al.*, “Radiomics Analysis at PET/CT Contributes to Prognosis of Recurrence and Survival in Lung Cancer Treated with Stereotactic Body Radiotherapy,” *Scientific Reports*, vol. 8, no. 1, 2018.

[20] H.J. Aerts, E.R. Velazquez, *et al.*, “Decoding Tumour Phenotype by Noninvasive Imaging Using a Quantitative Radiomics Approach,” *Nature Communications*, vol. 5, 2014.

[21] W. Sun, B. Zheng, *et al.*, “Automatic Feature Learning using Multichannel ROI based on Deep Structured Algorithms for Computerized Lung Cancer Diagnosis,” *Computers in Biology and Medicine*, vol. 89, no. 1, pp. 530-539, 2017.

[22] O. Echaniz, M. Grana, *et al.*, "Ongoing Work on Deep Learning for Lung Cancer Prediction," *Biomedical Applications Based on Natural and Artificial Computing*, vol. 10338, pp. 42-48, 2017.

[23] J. Griethuysen, A. Fedorov, *et al.*, "Computational Radiomics System to Decode the Radiographic Phenotype," *Cancer Research*, vol. 77, no. 21, pp. 104-107, 2017.

[24] Y. Zhang, A. Oikonomou, *et al.*, "Radiomics-based Prognosis Analysis for Non-small Cell Lung Cancer," *Scientific Reports*, vol. 7, 2017.

[25] W. Sun, B. Zheng, *et al.*, "Computer Aided Lung Cancer Diagnosis with Deep Learning Algorithms," *Proc.SPIE*, vol. 9785, 2016.

[26] H. Li, Y. Zhu, *et al.*, "MR Imaging Radiomics Signatures for Predicting the Risk of Breast Cancer Recurrence as Given by Research Versions of MammaPrint, Oncotype DX, and PAM50 Gene Assays.," *Radiology*, vol. 281, no. 2, pp. 382-391, 2016.

[27] A. Cunliffe, S. G.Armato III, *et al.*, "Lung Texture in Serial Thoracic Computed Tomography Scans: Correlation of Radiomics-based Features With Radiation Therapy Dose and Radiation Pneumonitis Development," *International Journal of Radiation Oncology*Biology*Physics*, vol. 91, no. 5, pp. 1048-1056, 2015.

[28] R. Berenguer, M.D.R. Pastor-Juan, *et al.*, "Radiomics of CT Features May Be Nonreproducible and Redundant: Influence of CT Acquisition Parameters," *Radiology*, vol. 288, no. 2, pp. 407-415, 2018.

[29] R. Thawani, M. McLane, *et al.*, "Radiomics and Radiogenomics in Lung Cancer: A Review for the Clinician," *Lung cancer*, vol. 115, pp. 34-41, 2017.

[30] M. Sasaki, K. Yamada, *et al.*, "Variability in Absolute Apparent Diffusion Coefficient Values Across Different Platforms may be Substantial: A Multivendor, Multi-institutional Comparison Study," *Radiology*, vol. 249, no. 2, pp. 624-630, 2008.

[31] M. Kolossvary, J. Karady, *et al.*, "Radiomic Features Are Superior to Conventional Quantitative Computed Tomographic Metrics to Identify Coronary Plaques With Napkin-Ring Sign," *Circulation: Cardiovascular Imaging*, vol. 10, no. 12, 2017.

[32] H. Suk, S. Lee, *et al.* " Hierarchical Feature Representation and Multimodal Fusion with Deep Learning for AD/MCI Diagnosis," *Neuroimage*, vol. 101, 2014.

[33] A. Rahmim, P. Huang, *et al.*, "Improved Prediction of Outcome in Parkinson's Disease using Radiomics Analysis of Longitudinal DAT SPECT Images," *NeuroImage: Clinical*, vol. 16, pp. 539-544, 2017.

[34] J. Dehmeshki, H. Amin, M. Valdivieso, X. Ye, "Segmentation of Pulmonary Nodules in Thoracic CT Scans: A Region Growing Approach," *IEEE Trans. Med. Imag*, vol. 27, no. 4, pp. 467-480, 2008.

[35] A. A. Farag, H. E. Abd El Munim, J. H. Graham, A. A. Farag, "A Novel Approach for Lung Nodules Segmentation in Chest CT Using Level Sets," *IEEE Trans. Image Process.*, vol. 22, no. 12, pp. 5202-5213, 2013.

[36] O. Ronneberger, Ph. Fischer, T. Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation," *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 234-241, 2015.

[37] M. Anthimopoulos, S. Christodoulidis, L. Ebner, T. Geiser, A. Christe, S. Mougiakakou, “Semantic Segmentation of Pathological Lung Tissue with Dilated Fully Convolutional Networks,” *arXiv*, preprint arXiv: 1803.06167, 2018.

[38] G. Huang, Z. Liu, L. van der Maaten, K. Q. Weinberger, “Densely Connected Convolutional Networks” *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[39] P. Wang, P. Chen, Y. Yuan, D. Liu, Z. Huang, X. Hou, G. Cottrell, ‘Understanding Convolution for Semantic Segmentation” *arXiv*, preprint arXiv: 1702.08502, 2018.

[40] V. Parekh, M. Jacobs, “Radiomics: A New Application from Established Techniques,” *Expert Review of Precision Medicine and Drug Development*, vol. 1, no. 2, pp. 207-226, 2016.

[41] M.J. Shafiee, A.G. Chung, *et al.*, “Discovery Radiomics via Evolutionary Deep Radiomic Sequencer Discovery for Pathologically Proven Lung Cancer Detection," *Journal of medical imaging*, vol. 4, no. 4, 2017.

[42] D. Kumar, A. Wong, *et al.*, “Lung Nodule Classification Using Deep Features in CT Images," *International Conference on Computer and Robot Vision*, 2015.

[43] J. Cheng, W. Yang, *et al.*, “Retrieval of Brain Tumors by Adaptive Spatial Pooling and Fisher Vector Representation,” *PloS one.*, 2016.

[44] F. Ciompi, K. Chung, *et al.*, “ Towards Automatic Pulmonary Nodule Management in Lung Cancer Screening with Deep Learning,” *Scientific Reports*, vol. 7, 2017.

[45] W. Shen, M. Zhou, *et al.*, “ Learning from Experts: Developing Transferable Deep Features for Patient-Level Lung Cancer Prediction,” *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp, 124-131, 2016.

[46] D. Kumar, A. Chung, *et al.*, “ Discovery Radiomics for Pathologically-Proven Computed Tomography Lung Cancer Prediction,” *Image Analysis and Recognition*, pp. 54-62, 2017.

[47] D. Ravi, C. Wong, *et al.*, “Deep Learning for Health Informatics," *IEEE Journal of Biomedical and Health Informatics*, vol. 21, no. 1, pp. 4-21, 2017.

[48] A. Jamaludin, T. Kadir, *et al.*, “ SpineNet: Automatically Pinpointing Classification Evidence in Spinal MRIs,” *International Conference on Medical Image Computing and Computer-Assisted Intervention*, vol. 9901, pp. 166-175, 2016.

[49] B. Huynh, H. Li, *et al.*, “ Digital Mammographic Tumor Classification Using Transfer Learning from Deep Vonvolutional Neural Networks,” *Journal of Medical Imaging*, vol. 3, no. 3, 2016.

[50] R. Paul, S. Hawkins, *et al.*, " Deep Feature Transfer Learning in Combination with Traditional Features Predicts Survival Among Patients with Lung Adenocarcinoma," *Tomography: a journal for imaging research*, vol. 2, no. 4, pp. 388-395, 2016.

[51] S. Sabour, N. Frosst, G.E. Hinton, "Dynamic Routing Between Capsules," *Conference on Neural Information Processing Systems (NIPS)*, pp. 3859-3869, 2017.

[52] W. Shen, *et al.*, " Multi-crop Convolutional Neural Networks for Lung Nodule Malignancy Suspiciousness Classification," *Pattern Recognition*, vol. 61, pp. 663-673, 2017.

[53] H. Wang, Z. Zhou, *et al.*, " Comparison of Machine Learning Methods for Classifying Mediastinal Lymph Node Metastasis of Non-Small Cell Lung Cancer from 18F-FDG PET/CT Images," *EJNMMI Research*, vol. 7, no. 1, 2017.

[54] C. Szegedy, V. Vanhoucke, *et al.*, "Rethinking the Inception Architecture for Computer Vision," *arXiv preprint arXiv:1512.00567v3*, 2015.

[55] C. Szegedy, W. Liu, *et al.*, "Going Deeper with Convolutions," *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[56] V. Gulshan, L. Peng, *et al.*, "Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs," *Journal of the American Medical Association (JAMA)*, vol. 16, no. 22, pp. 2402-2410, 2016.

[57] Z. Li, Y. Wang, *et al.*, " Deep Learning Based Radiomics (DLR) and Its Usage in Noninvasive IDH1 Prediction for Low Grade Glioma," *Scientific Reports*, vol. 7, no. 1, 2017.

[58] G. Litjens, T. Kooi, *et al.*, "A Survey on Deep Learning in Medical Image Analysis," *Medical Image Analysis*, vol. 42, pp. 60-88, 2017.

[59] L. Fu, J. Ma, *et al.*, " Automatic Detection of Lung Nodules: False Positive Reduction Using Convolutional Neural Networks and Handcrafted Features," *Proc.SPIE*, vol. 10134, 2017.

[60] L. Oakden-Rayner, G. Carneiro, *et al.*, " Precision Radiology: Predicting Longevity Using Feature Engineering and Deep Learning Methods in a Radiomics Framework," *Scientific Reports*, vol. 7, no. 1, 2017.

[61] S. Liu, H. Zhengr, *et al.*, " Prostate Cancer Diagnosis Using Deep Learning with 3D Multiparametric MRI," *Proc.SPIE*, vol. 10134, 2017.

[62] W. Shen, M. Zhou, *et al.*, " Multi-scale Convolutional Neural Networks for Lung Nodule Classification," *International Conference on Information Processing in Medical Imaging*, pp. 588-599, 2015.

[63] K. Liu, G. Kang, " Multiview Convolutional Neural Networks for Lung Nodule Classification," *International journal of imaging systems and technology*, vol. 27, no. 1, pp. 12-22, 2017.

[64] M. Liu, J. Zhang , *et al.*, " Anatomical Landmark based Deep Feature Representation for MR Images in Brain Disease Diagnosis," *IEEE Journal of Biomedical and Health Informatics*, 2018.

[65] S. Azizi, S.Bayat, *et al.*, "Deep Recurrent Neural Networks for Prostate Cancer Detection: Analysis of Temporal Enhanced Ultrasound," *IEEE Transactions on Medical Imaging*, 2018.

[66] B. Kim, Y. Sung, *et al.*, " Deep Feature Learning for Pulmonary Nodule Classification in a Lung CT," *2016 4th International Winter Conference on Brain-Computer Interface (BCI)*, 2016.

[67] N. Emaminejad, W. Qian, *et al.*, "Fusion of Quantitative Image and Genomic Biomarkers to Improve Prognosis Assessment of Early Stage Lung Cancer Patients," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 5, pp. 1034-1043, 2016.

[68] J. Lao, Y. Chen, *et al.*, " A Deep Learning-Based Radiomics Model for Prediction of Survival in Glioblastoma Multiforme," *Scientific Reports*, vol. 7, no. 1, 2017.

[69] J. Peng, J. Zhang, *et al.*, " A Radiomics Nomogram for Preoperative Prediction of Microvascular Invasion Risk in Hepatitis B Virus-Related Hepatocellular Carcinoma," *Diagnostic and Interventional Radiology*, vol. 24, no. 3, pp. 121-127, 2018.

[70] C. Shen, Z. Liu*et al.*, "Building CT Radiomics Based Nomogram for Preoperative Esophageal Cancer Patients Lymph Node Metastasis Prediction," *Translational Oncology*, vol. 11, no. 3, pp. 815-824, 2018.

[71] N. Antropova, B. Huynh, *et al.*, " A Deep Feature Fusion Methodology for Breast Cancer Diagnosis Demonstrated on Three Imaging Modality Datasets," *The International Journal of Medical Physics Research and Practice*, vol. 44, no. 10, pp. 5162-5171, 2017.

[72] S. Liu, Y. Xie, *et al.*, " Pulmonary Nodule Classification in Lung Cancer Screening with Three-Dimensional Convolutional Neural Networks," *Journal of Medical Imaging*, vol. 4, no. 4, 2017.

[73] S. Chen, J. Qin, *et al.*, " Automatic Scoring of Multiple Semantic Attributes With Multi-Task Feature Leverage: A Study on Pulmonary Nodules in CT Images," *IEEE Transactions on medical imaging*, vol. 36, no. 3, pp. 802-814, 2017.

[74] R. L. Siegel, K. D. Miller, A. Jemal, "Cancer Statistics," *A cancer journal for clinicians*, 2016.

[75] "Tumer Types: Understanding Brain Tumors," *National Brain Tumor Society*, 2018.

[76] J. Cheng, W. Huang, *et al.*, "Enhanced Performance of Brain Tumor Classification via Tumor Region Augmentation and Partition," *PloS one.*, 2015.

[77] M. Havaei, A. Davy, D. Warde-Farley, A. Biard, A. Courville, Y. Bengio, Ch. Palc, P. Jodoin, H. Larochelle, "Brain tumor segmentation with Deep Neural Networks," *Medical Image Analysis*, vol. 35, pp. 18-31, 2017.

[78] Kh. Usman, K. Rajpoot, "Brain tumor classification from multi-modality MRI using wavelets and machine learning," *Pattern Analysis and Applications*, vol. 20, no. 3, pp. 871-881, 2017.

[79] N. K. El Abbadi, N. E. Kadhim, "Brain Cancer classification Based on Features and Artificial Neural Network," *International Journal of Advanced Research in Computer and Communication Engineering*,vol. 8, no. 1, 2017.

[80] A. Krizhevsky, I. Sutskever, "ImageNet Classification with Deep Convolutional Neural Networks," *Neural Information Processing Systems (NIPS)*, 2012.

[81] J. S. Paul, A. J. Plassard, B. A. Landman, D. Fabbribi, "Deep Learning for Brain Tumor Classification," *PROCEEDINGS OF SPIE*, 2017.

[82] M. Moghimi, S.J. Belongie, *et al.*, "Boosted Convolutional Neural Networks.," *BMVC*, 2016.

[83] J. Zhu, S. Rosset, *et al.*, "Multi-class AdaBoost," *Statistics and its interface*, vol. 2, no. 3, 2006.

[84] H. Schwenk, Y. benjio, "Boosting Neural Networks," *Neural Computation*, vol. 12, no. 8, 2000.

[85] Y. Gal, Z. Ghahramani, "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning," *Proceedings of Machine Learning Research*, vol. 48, 2016.

[86] A. Kendall, Y. Gal, "What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?," *31st Conference on Neural Information Processing Systems (NIPS 2017)*, 2017.

[87] H. Wang, D. Yeung, "Towards Bayesian Deep Learning: A Survey," *ArXiv*, 2017.

[88] H. Wang, D. Yeung, "Towards Bayesian Deep Learning: A Framework and Some Existing Methods," *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*, vol. 28, no. 12, pp. 3395-3408, 2016.

[89] S. Amiri, M. Mahjoub, I. Rekik, "Bayesian Network and Structured Random Forest Cooperative Deep Learning for Automatic Multi-label Brain Tumor Segmentation," *Proceedings of the 10th International Conference on Agents and Artificial Intelligence (ICAART 2018)* , vol. 2, pp. 183-190, 2018.

[90] T. Nair, D. Precup, *et al.*, "Exploring uncertainty measures in deep networks for Multiple sclerosis lesion detection and segmentation," *Medical Image Analysis*, vol. 59, 2020.

[91] Z. Eaton-Rosen, F. Bragman, *et al.*, "Towards Safe Deep Learning: Accurately Quantifying Biomarker Uncertainty in Neural Network Predictions," *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 691-699, 2018.

[92] O. Ozdemir, B. Woodward, A.A. Berlin, "Propagating Uncertainty in Multi-Stage Bayesian Convolutional Neural Networks with Application to Pulmonary Nodule Detection," *Second workshop on Bayesian Deep Learning (NIPS 2017)*, 2017.

[93] F.S. Ribeiro, G. Leontidis, S. Kollias, "Capsule Routing via Variational Bayes," *The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)*, 2020.

[94] I. Ramirez, A. Cuesta-Infante, E. Schiavi, J.J. Pantrigo, "Bayesian capsule networks for 3D human pose estimation from single 2D images," *Neurocomputing*, vol. 379, pp. 64-73, 2020.

[95] A. Assa, K. N. Plataniotis, "Wasserstein-Distance-Based Gaussian Mixture Reduction," *IEEE Signal Processing Letters*, vol. 25, no. 10, pp. 1465-1469, 2018,

[96] A. Mohammadi, K. N. Plataniotis, "Improper Complex-Valued Bhattacharyya Distance," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, no. 5, pp. 1049-1064, 2016.

[97] A. Kendall, V. Badrinarayanan R. Cipolla, "Bayesian SegNet: Model Uncertainty in Deep Convolutional Encoder-Decoder Architectures for Scene Understanding," *Proceedings of the British Machine Vision Conference (BMVC)*, pp. 57.1-57.12, 2017.

[98] E. Hullermeier, W. Waegeman, "Aleatoric and Epistemic Uncertainty in Machine Learning: An Introduction to Concepts and Methods," *arXiv:1910.09457v3*, 2020.

[99] A. K. Ganti, J. L. Mulshine, "Lung cancer screening", *The Oncologist*, vol. 11, pp. 481-487, 2006.

[100] J. L. Causey, *et al.*, "Highly accurate model for prediction of lung nodule malignancy with ct scans.", *Scientific Reports*, vol. 8, 2018.

[101] S. G. Armato III, *et al.*, "Data from LIDC-IDRI.", *The Cancer Imaging Archive.*, 2015.

[102] S. G. Armato III, *et al.*, "The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): A Completed Reference Database of Lung Nodules on CT Scans.", *Medical Physics*, vol. 38, pp. 915-931, 2011.

[103] K. Clark, *et al.*, "The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository", *Journal of Digital Imaging*, vol. 26, pp. 1045-1057, 2013.

[104] A. G. Lalkhen, A. McCluskey, "Non-Small Cell Lung Cancer: Epidemiology, Risk Factors, Treatment, and Survivorship", *Continuing Education in Anaesthesia Critical Care and Pain*, vol. 8, pp. 221-223, 2008.

[105] T. Brosch, *et al.*, Deep convolutional encoder networks for multiple sclerosis lesion segmentation", *Medical Image Computing and Computer-Assisted Intervention, MICCAI 2015. MICCAI 2015. Lecture Notes in Computer Science, Springer, Cham*, vol. 9351, 2015.

[106] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, M. Jorge Cardoso, “Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations", *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. DLMIA 2017, ML-CDS 2017. Lecture Notes in Computer Science, Springer, Cham*, vol. 10553, 2017.

[107] C. Jacobs, *et al.*, “Automatic detection of subsolid pulmonary nodules in thoracic computed tomography images", *Medical Image Analysis*, vol. 18, pp. 374-384, 2014.

[108] D. Kumar, *et al.*, “Discovery radiomics for pathologically-proven computed tomography lung cancer prediction", *Karray F., Campilho A., Cheriet F. (eds) Image Analysis and Recognition. ICIAR 2017. Lecture Notes in Computer Science, Springer, Cham*, vol. 10317, 2017.

[109] K. Liu, G. Kang, “Multiview convolutional neural networks for lung nodule classification", *International Journal of Imaging Systems and Technology*, vol. 27, pp. 12-22, 2017.

[110] A. Nibali, H. Zhen, D. Wollersheim, “Pulmonary nodule classification with deep residual networks", *International Journal of Computer Assisted Radiology and Surgery*, vol. 12, pp. 1799-1808, 2017.

[111] Y. Xie, J. Zhang, S. Liu, W. Cai, Y. Xia, “Lung nodule classification by jointly using visual descriptors and deep features", *Medical Computer Vision and Bayesian and Graphical Models for Biomedical Imaging. BAMBI 2016, MCV 2016. Lecture Notes in Computer Science, Springer, Cham*, vol. 10081, 2017.

[112] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, G. E. Hinton, “Adaptive Mixtures of Local Experts,” *Neural Computation*, vol. 3, pp. 79-87, 1991.

[113] R. Rasti, M. Teshnehlab, S. M. Phung, “Breast Cancer Diagnosis in DCE-MRI Using Mixture Ensemble of Convolutional Neural Networks,” *Pattern Recognition*, vol. 72, pp. 381-390, 2017.

[114] J. Guo, S. Gould, “Deep CNN Ensemble with Data Augmentation for Object Detection,” *arXiv:1506.07224*, 2015.

[115] D. Maji, A. Santara, P. Mitra, D. Sheet, “Ensemble of Deep Convolutional Neural Networks for Learning to Detect Retinal Vessels in Fundus Images,” *arXiv:1603.04833*, 2016.

[116] H. Chen, Q. Dou, X. Wang, J. Qin, P. Heng, “Mitosis Detection in Breast Cancer Histology Images via Deep Cascaded Networks,” *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pp. 1160-1166, 2016.

[117] H. Wang, A. Cruz-Roa, A. Basavanhally, H. Gilmore, N. Shih, M. Feldman, J. Tomaszewski, F. Gonzalez, A. Madabhushi, “Mitosis Detection in Breast Cancer Pathology Images by Combining Handcrafted and Convolutional Neural Network Features,” *Journal of Medical Imaging*, vol. 1, 2014.

[118] M. Jordan, R. Jacobes, “Hierarchical Mixtures of Experts and the EM Algorithm,” *Proceedings of 1993 International Joint Conference on Neural Networks*, pp. 1339-1344, 1993.

[119] T. Hahn, M. Pyeon G. Kim, “Self-Routing Capsule Networks,” *advances in Neural Information Processing Systems*, pp. 7658-7667, 2019.

[120] A. Mohammadi, K. Plataniotis, “Improper Complex-Valued Multiple-Model Adaptive Estimation,” *IEEE Transactions on Signal Processing*, vol. 63, pp. 1528-1542, 2015.

[121] G. Hinton, V. Oriol, J. Dean, “Distilling the Knowledge in a Neural Network,” *NIPS Deep Learning and Representation Learning Workshop*, 2015.

[122] L. R. Pilz, C. Manegold, G. Schmid-Bindert, “Statistical considerations and endpoints for clinical lung cancer studies: can progression free survival (pfs) substitute overall survival (os) as a valid endpoint in clinical trials for advanced non-small-cell lung cancer?", *Translational Lung Cancer Research*, vol. 1, 2012.

[123] A. Arnett, *et al.*, “Long-term clinical outcomes and safety profile of sbrt for centrally located nsclc.", *Adv Radiat Oncol*, vol. 4, pp. 422-428, 2019.

[124] J. Wu, *et al.*, “Early-stage non–small cell lung cancer: Quantitative imaging characteristics of 18f fluorodeoxyglucose pet/ct allow prediction of distant metastasis", *Radiology*, vol. 281, pp. 270-278, 2016.

[125] T. Pyka, *et al.*, “Textural features in pre-treatment [f18]-fdg-pet/ct are correlated with risk of local recurrence and disease-specific survival in early stage nsclc patients receiving primary stereotactic radiation therapy", *Radiation Oncology*, vol. 10, 2015.

[126] Y. Huang, *et al.*, “Radiomics signature: A potential biomarker for the prediction of disease-free survival in early-stage (i or ii) non—small cell lung cancer", *Radiology*, vol. 281, pp. 947-957, 2016.

[127] S. Leger, *et al.*, “A comparative study of machine learning methods for time-to-event survival data for radiomics risk modelling", *Scientific Reports*, vol. 7, 2017.

[128] R. J. Gillies, P. E. Kinahan, H. Hricak, “Radiomics: Images are more than pictures, they are data", *Radiology*, 2015.

[129] C. Chen, *et al.*, “Radiomic features analysis in computed tomography images of lung nodule classification.", *PLoS One*, vol. 13, 2018.

[130] S. S. F. Yip, H. J. W. L. Aerts, “Applications and limitations of radiomics", *Physics in Medicine and Biology*, vol. 61, 2016.

[131] C. Parmar, *et al.*, “Radiomic feature clusters and prognostic signatures specific for lung and head and neck cancer", *Scientific Reports*, vol. 5, 2015.

[132] W. Sun, M. Jiang, J. Dang, P. Chang, F. Yin, “Effect of machine learning methods on predicting nsclc overall survival time based on radiomics analysis", *Radiation Oncology*, vol. 13, 2018.

[133] J. E. Timmeren, *et al.*, “Survival prediction of non-small cell lung cancer patients using radiomics analyses of cone-beam ct images", *Radiotherapy and Oncology*, vol. 123, 2017.

[134] M. Khorrami, *et al.*, “Combination of peri- and intratumoral radiomic features on baseline ct scans predicts response to chemotherapy in lung adenocarcinoma", *Radiology: Artificial Intelligence*, vol. 1, pp. 180012, 2019.

[135] M. Pavic, *et al.*, “Influence of inter-observer delineation variability on radiomics stability in different tumor sites", *Acta Oncologica*, vol. 57, pp. 1070-1074, 2018.

[136] H. Li, *et al.*, “Deep convolutional neural networks for imaging data based survival analysis of rectal cancer", *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pp. 846-849, 2019.

[137] A. Krizhevsky, I. Sutskever, G. E. Hinton, “Imagenet classification with deep convolutional neural networks", *Neural Information Processing Systems (NIPS)*, 2012.

[138] R. Yamashita, M. Nishio, R. K. G. Do, K. Togashi, “Convolutional neural networks: An overview and application in radiology", *Insights into Imaging*, vol. 9, pp. 611–629, 2018.

[139] X. Zhu, J. Yao, J. Huang, “Deep convolutional neural network for survival analysis with pathological images", *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 544-547, 2016.

[140] A. Kandathil, F. U. Kay, Y. M. Butt, L. W. Wachsmann, R. M. Subramaniam, “Role of fdg pet/ct in the eighth edition of tnm staging of non–small cell lung cancer", *Nuclear Medicine*, vol. 38, pp. 2134–2149, 2018.

[141] Y. Xu, *et al.*, “Deep learning predicts lung cancer treatment response from serial medical imaging", *Precision Medicine and Imaging*, 2019.

[142] S. Yoo, *et al.*, “Prostate Cancer Detection using Deep Convolutional Neural Networks", *Scientific Reports*, vol. 9, 2019.

[143] R. Dey, L. Zh, Y. Hong, “Diagnostic classification of lung nodules using 3d neural networks", *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, 2018.

[144] R. Paul, *et al.*, “Deep feature transfer learning in combination with traditional features predicts survival among patients with lung adenocarcinoma", *Tomography* vol. 2, pp. 388-395, 2016.

[145] E. Huynh, *et al.*, “Ct-based radiomic analysis of stereotactic body radiation therapy patients with lung cancer. *Radiotherapy and Oncology"*, vol. 120, pp. 258-266, 2016.

[146] H. Wang, *et al.*, “Comparison of machine learning methods for classifying mediastinal lymph node metastasis of non-small cell lung cancer from 18f-fdg pet/ct images", *EJNMMI Research*, vol. 7, 2017.

[147] M. Luck, T. Sylvain, H. Cardinal, A. Lodi, Y. Bengio, “Deep learning for patient-specific kidney graft survival analysis", *arXiv e-prints*, 2017.

[148] V. Cheplygina, M. Bruijne, J. P. W. Pluim, “Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis", *Medical Image Analysis*, vol. 54, pp. 280 -296, 2019.

[149] P. Vincent, H. Larochelle, Y. Bengio, P. Manzagol, “Extracting and composing robust features with denoising autoencoders", *Proceedings of the 25th International Conference on Machine Learning*, pp. 1096-1103, 2008.

[150] J. Masci, U. Meier, D. Ciresan, J. Schmidhuber, “Stacked convolutional auto-encoders for hierarchical feature extraction", *Artificial Neural Networks and Machine Learning – ICANN 2011*, pp. 52-59, 2011.

[151] R. Al Umairi, *et al.*, “Can radiomics at ct and staging pet/ct serve as an imaging biomarker of the egfr mutation and alk fusion in lung adenocarcinoma?", *J Thorac Imaging*, vol. 32, 2017.

[152] H. Steck, B. Krishnapuram, C. Dehing-oberije, P. Lambin, V. C. Raykar, “On ranking in survival analysis: Bounds on the concordance index", *Advances in Neural Information Processing Systems 20*, pp. 1209-1216, 2008.

[153] R. Thawani, *et al.*, “Radiomics and radiogenomics in lung cancer: A review for the clinician", *Lung Cancer*, vol. 115, 34-41, 2018.

[154] H. Ishwaran, *et al.*, “Random Survival Forests", *Ann. Appl. Stat.*, vol. 2, pp. 841-860, 2008.

[155] H. Steck, *et al.*, “On Ranking in Survival Analysis: Bounds on the Concordance Index", *Advances in Neural Information Processing Systems 20 (NIPS)* , 2007.

[156] J. Kang, *et al.*, “Predicting 5-Year Progression and Survival Outcomes for Early Stage Non-small Cell Lung Cancer Treated with Stereotactic Ablative Radiation Therapy: Development and Validation of Robust Prognostic Nomograms", *Radiation Oncology*, vol. 106, 90-99, 2020.

[157] M. Schmid, M.N. Wright, A. Ziegler, “On the use of Harrell's C for clinical risk prediction via random survival forests", *Expert Systems with Applications: An International Journal*, vol. 63, 2016.

[158] M. K. Goel, *et al.*, “Understanding Survival Analysis: Kaplan-Meier Estimate", *International Journal of Ayurveda Research*, vol. 1, pp. 274-278, 2010.

[159] J. M. Bland, *et al.*, “The logrank test", *BMJ*, vol. 328, 2004.

[160] S. Wang, *et al.*, “Unsupervised deep learning features for lung cancer overall survival analysis.", *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2018.

[161] J. B. Nasejje, H. Mwambi, K. Dheda, M. Lesosky, “A comparison of the conditional inference survival forest model to random survival forests based on a simulation study as well as on two applications with time-to-event data", *BMC Medical Research Methodology*, vol. 17, 2017.

[162] X. Xu, X. Jiang, *et al.*, “A Deep Learning System to Screen Novel Coronavirus Disease 2019 Pneumonia,” *Engineering*, 2020.

[163] Sh. Wang, B. Kang, *et al.*, “A Deep Learning Algorithm using CT Images to Screen for Corona Virus Disease (COVID-19),” *medRxiv*, 2020.

[164] O. Gozes, M. Frid-Adar, *et al.*, “Rapid AI Development Cycle for the Coronavirus (COVID-19) Pandemic: Initial Results for Automated Detection & Patient Monitoring using Deep Learning CT Image Analysis ,” *arXiv:2003.05037*, 2020

[165] A. Narin, C. Kaya, Z. Pamuk , “Automatic Detection of Coronavirus Disease (COVID-19) Using X-ray Images and Deep Convolutional Neural Networks ,” *arXiv:2003.10849*, 2020

[166] M. Farooq, A. Hafeez, “COVID-ResNet: A Deep Learning Framework for Screening of COVID19 from Radiograph,” *arXiv:2003.14395*, 2020

[167] R. Yamashita, M. Nishio, *et al.*, “Convolutional Neural Networks: An Overview and Application in Radiology,” *Insights into Imaging*, vol. 9, no. 4, pp. 611-629, 2018.

[168] L. Li, L. Qin, *et al.*, “Artificial Intelligence Distinguishes COVID-19 from Community Acquired Pneumonia on Chest CT,” *Radiology*, 2020

[169] L. Wang, A. Wong, “COVID-Net: A Tailored Deep Convolutional Neural Network Design for Detection of COVID-19 Cases from Chest Radiography Images,” *arXiv:2003.09871*, 2020.

[170] P. K. Sethy, S. K. Behera, “Detection of Coronavirus Disease (COVID-19) Based on Deep Features,” *Preprints 2020, 2020030300*, 2020.

[171] X. Wang, Y. Peng, *et al.*, “ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases,” *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3462-3471, 2017.

[172] J. P. Cohen, “COVID Chest x-ray Dataset,” *https://github.com/ieee8023/covid-chestxray-dataset*, 2020.

[173] P. Mooney, “Kaggle Chest x-ray Images (Pneumonia) Dataset,” *https://github.com/ieee8023/covid-chestxray-dataset*, 2020.

[174] B. Xu, Y. Xing, *et al.*, “Chest CT for detecting COVID-19: a systematic review and meta-analysis of diagnostic accuracy,” *European Radiology*, pp. 1-8, 2020.

[175] V. Sze, Y. Chen, *et al.*, “Efficient Processing of Deep Neural Networks: A Tutorial and Survey,” *Proceedings of the IEEE*, vol. 105, no. 12, pp. 2295-2329, 2017.

[176] A. Borakati, A. Perera, J. Johnson, T. Sood, “Diagnostic accuracy of X-ray versus CT in COVID-19: a propensity-matched database study", *BMJ Open*, vol. 10, pp. e042946, 2020.

[177] Z. Sun, N. Zhang, Y. Li, X. Xu, “A systematic review of chest imaging findings in covid-19", *Quantitative imaging in medicine and surgery*, vol. 10, pp. 1058-1079, 2020.

[178] H. Bjorke, “Covid-19 segmentation dataset", *MedSeg*, 2020.

[179] M. Jun, *et al.*, “Covid-19 ct lung and infection segmentation dataset (version 1.0"), *Zenodo*, 2020.

[180] J. P. Cohen, P. Morrison, L. Dao, “Covid-19 image data collection", Preprint at `http://arxiv.org/abs/2003.11597`, 2020.

[181] S. Morozov, *et al.*, “Mosmeddata: Chest ct scans with covid-19 related findings dataset", Preprint at `https://arxiv.org/abs/2005.06465`, 2020.

[182] J. Zhao, Y. Zhang, X. He, P. Xie, “Covid-ct-dataset: a ct scan dataset about covid-19", Preprint at `https://arxiv.org/abs/2003.13865`, 2020.

[183] E. Soares, P. Angelov, S. Biaso, M. Higa Froes, D. Kanda Abe, “Sars-cov-2 ct-scan dataset: A large dataset of real patients ct scans for sars-cov-2 identification", Preprint at medRxiv, 2020.

[184] M. Rahimzadeh, A. Attar, S. Sakhaei, “A fully automated deep learning-based network for detecting covid-19 from a new and large lung ct scan dataset", Preprint at medRxiv, 2020.

[185] J. Jacob, *et al.*, “Using imaging to combat a pandemic: rationale for developing the uk national covid-19 chest imaging database", *European Respiratory Journal*, vol. 56, pp. 2001809, 2020.

[186] A. P. Dhawan, “Medical Imaging Modalities: X-Ray Imaging", *Medical Image Analysis*, pp. 79-97, 2011.

[187] A. Katsevich, “Theoretically Exact Filtered Backprojection-Type Inversion Algorithm for Spiral CT", *SIAM Journal on Applied Mathematics*, vol. 62, pp. 2012-2026, 2002.

[188] E. Seeram, “Computed Tomography: Physical Principles and Recent Technical Advances", *Journal of Medical Imaging and Radiation Sciences*, vol. 41, pp. 87-109, 2010.

[189] A. Bhalla, *et al.*, “Imaging protocols for ct chest: A recommendation", *Indian Journal of radiology and imaging*, vol. 29, pp. 236-246, 2019.

[190] D. S. Committee, W. Group, C. Trials, F. Text, “Supplement 142: Clinical Trial De-identification Profiles", *DICOM Standard*, pp. 1-44, 2011.

[191] M. Francone, *et al.*, “Chest ct score in covid-19 patients: correlation with disease severity and short-term prognosis", *European Radiology*, vol. 30, pp. 6808-6817, 2020.

[192] G. Bwire, “Coronavirus: Why men are more vulnerable to covid-19 than women?", *SN Comprehensive Clinical Medicine*, vol. 2, pp. 874-876, 2020.

[193] P. Afshar, *et al.*, “Covid-ct-md: Covid-19 computed tomography (ct) scan dataset applicable in machine learning and deep learning" *Figshare*, `https://figshare.com/s/5f297ea284f259621dce`, 2021.

[194] T. Ozturk, *et al.*, “Automated detection of covid-19 cases using deep neural networks with x-ray images", *Comput Biol Med*, 2020.

[195] T. Yan, *et al.*, “Automatic distinction between covid-19 and common pneumonia using multi-scale convolutional neural network on chest ct scans", *Chaos Solitons Fractals*, 2020.

[196] D. Fan, *et al.*, “Inf-Net: Automatic COVID-19 Lung Infection Segmentation From CT Images. *IEEE Transactions on Medical Imaging"*, vol. 39, pp. 2626-2637, 2020.

[197] Y. Mirsky, T. Mahler, I. Shelef, Y. Elovici, “Ct-gan: Malicious tampering of 3d medical imagery using deep learning", In *28th USENIX Security Symposium (USENIX Security 19)*, pp. 461-478, 2019.

[198] J. Hofmanninger, F. Prayer, J. Pan, S. Rohrich, H. Prosch, and G. Langs, “Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem,” 2020.

[199] K. Zhang, *et al.*, “Clinically Applicable AI System for Accurate Diagnosis, Quantitative Measurements, and Prognosis of COVID-19 Pneumonia Using Computed Tomography,” *Cell*, vol. 181, no. 6, pp. 1423-1433.e11, 2020.

[200] S. Hu, *et al.*, “Weakly Supervised Deep Learning for COVID-19 Infection Detection and Classification From CT Images,” *IEEE Access*, vol. 8, pp. 118869-118883, 2020.

[201] A. Criminisi, J. Shotton, E. Konukoglu, “ Decision Forests: A Unified Framework for Classification, Regression, Density Estimation, Manifold Learning and Semi-Supervised Learning", vol. 7, *NOW Publishers, foundations and trends® in computer graphics and vision*, vol. 7, no. 2-3, pp. 81-227, 2012.

[202] L. Li, *et al.*, “Using Artificial Intelligence to Detect COVID-19 and Community-acquired Pneumonia Based on Pulmonary CT: Evaluation of the Diagnostic Accuracy,” *Radiology*, vol. 296, no. 2, pp. E65-E71, 2020.

[203] Miglioretti, D. *et al.* The use of computed tomography in pediatrics and the associated radiation exposure and estimated cancer risk. *JAMA Pediatr* **167**, 700–707 (2013).

[204] Fearon, T. & Vucich, J. Pediatric patient exposures from ct examinations: Ge ct/t 9800 scanner. *AJR* **144**, 805–809 (1985).

[205] Pearce, M. *et al.* Radiation exposure from ct scans in childhood and subsequent risk of leukaemia and brain tumours: a retrospective cohort study. *Lancet* **380**, 499–505 (2012).

[206] Bell, D. & Gerstmair, A. As low as reasonably achievable (alara).

[207] Taekker, M., Kristjansdottir, B., Ole, G., Laursen, C. & Pietersen, P. Diagnostic accuracy of lowdose and ultra-low-dose ct in detection of chest pathology: a systematic review. *Clinical Imaging* **74**, 139–148 (2021).

[208] Sakane, H. *et al.* Biological effects of lowdose chest ct on chromosomal dna. *Radiology* **295**, 439–445 (2020).

[209] Park, J. *et al.* The usefulness of low-dose ct scan in elderly patients with suspected acute lower respiratory infection in the emergency room. *The British Journal of Radiology* **89** (2016).

[210] Schulze-Hagen, M. *et al.* Low-dose chest ct for the diagnosis of covid-19—a systematic, prospective comparison with pcr. *Dtsch Arztebl Int* **117**, 389–395 (2020).

[211] Dangis, A. *et al.* Accuracy and reproducibility of low-dose submillisievert chest ct for the diagnosis of covid-19. *Radiol Cardiothorac Imaging* **2**, e200196 (2020).

[212] Tofighi, S., Najafi, S., Johnston, S. & Gholamrezanezhad, A. Low-dose ct in covid-19 outbreak: radiation safety, image wisely, and image gently pledge. *Emergency Radiology* **27**, 601–605 (2020).

[213] Shiri, I. *et al.* Ultra-low-dose chest ct imaging of covid-19 patients using a deep residual neural network. *European Radiology* **31**, 1420–1431 (2021).

[214] Mei, X. *et al.* Artificial intelligence–enabled rapid diagnosis of patients with covid-19. *Nature Medicine* **26**, 1224–1228 (2020).

[215] Nilashi, M., Ibrahim, O. & Ahani, A. Accuracy improvement for predicting parkinson's disease progression. *Scientific Reports* **6** (2016).

[216] McNemar, Q. Psychological statistics. *Wiley* (1962).

[217] Prokop, M. *et al.* Co-rads: A categorical ct assessment scheme for patients suspected of having covid-19—definition and evaluation. *Radiology* **296**, E97–E104 (2020).

[218] Selvaraju, R. R. *et al.* Grad-cam: Visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (2017).

[219] Wu, E. *et al.* How medical ai devices are evaluated: limitations and recommendations from an analysis of fda approvals. *Nature Medicine* **27**, 576–584 (2021).

[220] Kleppe, A. *et al.* Designing deep learning studies in cancer diagnostics. *Nature Reviews Cancer* **21**, 199–211 (2021).

[221] Tan, M. *et al.* EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *Proceedings of the 36th International Conference on Machine Learning* **97**, 6105–6114 (2019).

[222] Tichavsky, P. *et al.* Posterior Cramer–Rao Bounds for ´ Discrete-Time Nonlinear Filtering. *IEEE Transactions on Signal Processing* **46**, 1386–1396 (1998).