Oğuzhan Ayyıldız

A Ph.D. Thesis

AGU 2022

MACHINE AND DEEP LEARNING BASED ANALYSIS OF TUMORS ON FDG-PET IMAGES

A THESIS
SUBMITTED TO THE DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING
AND THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE
OF ABDULLAH GUL UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

By
Oğuzhan Ayyıldız

June 2022

MACHINE AND DEEP LEARNING BASED ANALYSIS OF TUMORS ON FDG-PET IMAGES

A THESIS

SUBMITTED TO THE DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

AND THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE OF ABDULLAH GUL UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

By

Oğuzhan Ayyıldız

June 2022

SCIENTIFIC ETHICS COMPLIANCE

I hereby declare that all information in this document has been obtained in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all materials and results that are not original to this work.

Name-Surname: Oğuzhan Ayyıldız
Signature :

REGULATORY COMPLIANCE

Ph.D. thesis titled Machine and deep learning based analysis of tumors on FDG-PET images has been prepared in accordance with the Thesis Writing Guidelines of the Abdullah Gül University, Graduate School of Engineering & Science.

Prepared By
Oğuzhan Ayyıldız

Advisor
Prof. Dr. Bülent Yılmaz

Head of the Electrical and Computer Engineering Program
Assoc. Prof. Dr. Kutay İçöz

ACCEPTANCE AND APPROVAL

Ph.D. thesis titled Machine and deep learning based analysis of tumors on FDG-PET images prepared by Oğuzhan Ayyıldız has been accepted by the jury in the Electrical and Computer Engineering Graduate Program at Abdullah Gül University, Graduate School of Engineering & Science.

23 /06 / 2022

JURY:

Advisor : Prof. Bülent Yılmaz……………………………….....................................

Member : Prof. Semra İçer ……………………………..

Member : Assoc. Prof. Kutay İçöz……………………………………….......................

Member : Assoc. Prof. Seyhan Karaçavuş ………………………………………………

Member : Assist. Prof. Bekir Hakan Aksebzeci………………………………................

APPROVAL:

The acceptance of this Ph.D. thesis has been approved by the decision of the Abdullah Gül University, Graduate School of Engineering & Science, Executive Board, dated ….. /….. / ……….. and numbered ………….…….. .

……….. /……….. / ………..

(Date)

Graduate School Dean
Prof. Dr. İrfan ALAN

ABSTRACT

MACHINE AND DEEP LEARNING BASED ANALYSIS OF TUMORS ON FDG-PET IMAGES

Oğuzhan Ayyıldız
Ph.D. in Electrical and Computer Engineering
Advisor: Prof. Dr. Bülent Yılmaz

June 2022

Analysis of a tumor is essential in treatment planning and evaluation of treatment response. Positron Emission Tomography (PET) is a vital imaging device for clinical oncology in understanding the metabolic structure of the tumor. In this thesis, three separate studies investigating the application of machine, deep learning and statistical approaches on FDG-PET images from patients with non-small cell lung cancer (NSCLC) and pancreatic cancer. The first study aimed at performing a survey on subtype classification of NSCLC by using different texture features, feature selection methods and classifiers. Images from 92 patients and several clinical and metabolic features for each case were used in this study along with histopathological validation for the tumor subtype labeling. Stacking classifier resulted in 76% accuracy. The aim of our second study was to adapt an atrous (dilated) convolution-based tumor segmentation approach (DeepLabV3) on FDG-PET slices with maximum standard uptake value (SUVmax). MobileNet-v2 pretrained on ImageNet served as the backbone to DeepLabV3. The classification layer was interchanged with the Tversky loss layer which helped improve model's performance while the dataset was imbalanced. Images from 141 patients were employed and augmentation was performed in each training phase. Dice similarity index was obtained as 0.76 without preprocessing and 0.85 with preprocessing. The last study focused on determining the features to be used in the prognosis of pancreatic adenocarcinoma on FDG-PET images from 72 patients. Well-known texture, metabolic and physical features were extracted from tumor region that was determined with the help of random walk segmentation algorithm. On these features time-dependent ROC curve analysis was performed for 2-year overall survival (OS) prediction, and, in the univariable analyses, tumor size, energy, entropy, and strength were found to be significant predictors of OS.

Keywords: PET/CT, NSCLC, Machine learning, Deep learning, Radiomics, Semantic segmentation

ÖZET

FDG-PET GÖRÜNTÜLERİNDEKİ TÜMÖRLERİN MAKİNE VE DERİN ÖĞRENME TABANLI ANALİZİ

Oğuzhan Ayyıldız
Elektrik ve Bilgisayar Mühendisliği Anabilim Dalı Doktora
Tez Yöneticisi: Prof. Dr. Bülent Yılmaz

Haziran-2022

Bir tümörün analizi, tedavi planlamasında ve tedavi yanıtının değerlendirilmesinde esastır. Pozitron Emisyon Tomografisi (PET), tümörün metabolik yapısını anlamada klinik onkoloji için hayati bir görüntüleme cihazıdır. Bu tezde, küçük hücreli dışı akciğer kanseri (KHDAK) ve pankreas kanseri olan hastalardan alınan FDG PET görüntüleri üzerinde makine öğrenmesi, derin öğrenme ve istatistiksel yaklaşımların uygulanmasını araştıran üç ayrı çalışma yer almaktadır. İlk çalışma, farklı doku özellikleri, öznitelik seçim yöntemleri ve sınıflandırıcılar kullanılarak KHDAK'nin alt tip sınıflandırmasına odaklanmıştır. Bu çalışmada, tümör alt tipi etiketlemesi için histopatolojik doğrulama ile birlikte 92 hastanın görüntüleri ve her vaka için çeşitli klinik ve metabolik özellikler kullanılmıştır. İstifleme sınıflandırıcısı %76 doğrulukla sonuçlanmıştır. İkinci çalışmamızın amacı, maksimum standart alım değeri (SUVmax) bulunan FDG PET dilimlerinde atröz (dilate) evrişim tabanlı tümör segmentasyon yaklaşımını (DeepLabV3) uyarlamaktır. DeepLabV3'ün omurgası olarak ImageNet üzerinde önceden eğitilmiş MobileNet-v2 kullanılmıştır. Sınıflandırma katmanı, veri kümesi dengesizken modelin performansını iyileştirmeye yardımcı olan Tversky kayıp katmanıyla değiştirilmiştir. Her eğitim aşamasında 141 hastadan görüntüler ve büyütme kullanılmıştır. Dice benzerlik indeksi ön işleme olmadan 0,76 ve ön işleme ile 0,85 olarak elde edilmiştir. Son çalışma, 72 hastanın FDG PET görüntülerinde pankreas adenokarsinomunun prognozunda kullanılacak özelliklerin belirlenmesine odaklanmıştır. Rastgele yürüyüş segmentasyon algoritmasından yararlanarak elde edilen tümör bölgesinden, en sık kullanılan tekstür özellikleri, metabolk ve fiziksel özellikler çıkarılmış ve bu özellikler üzerinde, 2 yıllık genel sağkalım (GS) tahmini için zamana bağlı ROC eğrisi analizi gerçekleştirilmiştir. Tek değişkenli analizlerde, tümör boyutu, enerjisi, entropi ve gücü, GS'nin önemli belirleyicileri olarak tespit edilmiştir.

Keywords: PET/BT, KHDAK, Makine öğrenmesi, Derin öğrenme, Radyomiks, Semantik bölütleme

Acknowledgments

This thesis is supported by TÜBİTAK (The Scientific and Technological Research Council of Turkey) under Project No: 118E188.

First, I thank my advisor, Prof. Bülent Yılmaz. I am grateful to him for showing great patience and understanding during our work.

I would also like to thank Prof. Seyhan Karaçavuş. She always took care of me like a second advisor.

I would like to thank Prof. Kutay İçöz for his contributions throughout the thesis and his efforts to improve the thesis.

I would not have been able to complete this thesis without the support of my family. I want to express my gratitude to my dear mother and father, Enise and Nusret Ayyıldız, and my brother Burak Celal Ayyıldız, who raised me, supported me, and always listened to me with understanding.

Finally, I would like to thank my dear friends Doğuş Azer Kaçmaz and Fatih Altındiş. I am grateful for our pleasant conversations on a wide range of topics, from literature to technology, and for their unconditional support whenever I find myself in a tricky situation.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

PET	Positron Emission Imaging
CT	Computer Tomography
MRI	Magnetic Resonance Imaging
ML	Machine Learning
DL	Deep Learning
RW	Random Walk
VOI	Volume of Interest
ROI	Region of Interest
SUV	Standardized Uptake Value
NSCLC	Non-Small Cell Lung Cancer
ADC	Adenocarcinoma
SqCC	Squamous Cell Carcinoma
PA	Pancreatic adenocarcinoma
GLCM	Gray Level Co-Occurrence Matrix
GLRLM	Gray Level Run Length Matrix
NGTDM	Neighborhood Gray Tone Difference Matrix
GSZM	Gray Size Zone Matrix
MTV	Metabolic Tumor Volume
FDG	18F-fluoro-2-deoxy-D-glucose
ROC	Receiver Operating Characteristic
AUC	Area Under Curve
LBF	Local Binary Pattern
DICOM	Digital Imaging and Communications in Medicine
RBF	Radial Basis Function
SVM	Support Vector Machine
CFS	Correlation Based Feature Selection
k-NN	k-Nearest Neighbors
WEKA	Waikato Environment for Knowledge Analysis
LOOCV	Leave-One-Out Cross-Validation
MCC	Matthew’s Correlation Coefficient

PPV	Positive Predictive Value
NPV	Negative Predictive Value
DSI	Dice Similarity Index
SAMD	Software as a Medical Device
FDA	Food and Drug Administration

Chapter 1

Introduction

1.1 Positron Emission Tomography (PET) Imaging

1.1.1 Working Principle

Positron emission tomography (PET) is an imaging method in which we can observe cellular and molecular events. To monitor biological changes, we need tracers which are radiolabeled molecular probes. These tracers help measure cell proliferation, perfusion, oxygen metabolism, tumor-receptor density, and reporter-gene expression [1]. Frequently used isotopes are O-15, N-13, C-11, and F-18.

The radionuclide is injected into a vein. Then PET scanner moves to the part of the body examined. The annihilation of photons creates gamma rays. The PET camera detects coincident gamma rays emitted by the patient. Then images are reconstructed based on the related location and concentration of the tracer [1].

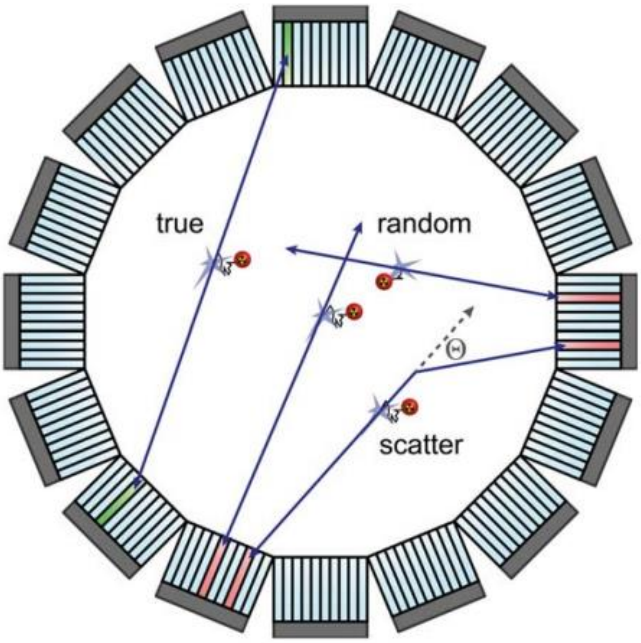


Figure 1. 1 PET ring schematic [2]

In Figure 1.1, showing the example of a PET ring, the image is reconstructed based on true coincidence. In the modern PET system, the scattering of photons is detected and eliminated before the reconstruction to improve image resolution.

The amount of radionuclide affects the brightness of the tissue. For instance, 2-18F-fluoro-2-deoxy-D-glucose (FDG) accumulates in cancer cells, making cancer cells brighter in images than healthy tissue. Figure 1.2 shows an FDG PET image of a lung cancer patient from our patient database.

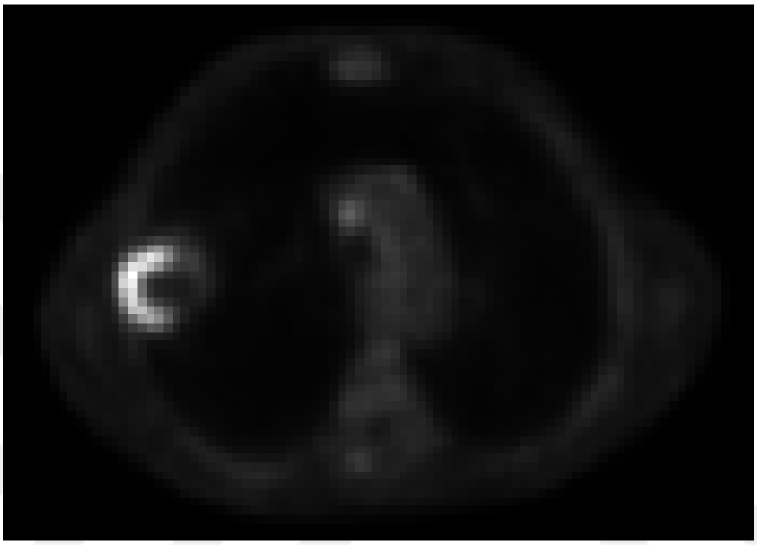


Figure 1. 2 Sample FDG PET image from our database

1.1.2 FDG PET/CT in oncology

Standardized uptake value (SUV) is a semi-quantitative assessment value used in PET images to show the metabolic activity of a tumor. SUV is calculated as a ratio of FDG concentration on the region of interest (ROI) to the injected dose normalized to patient weight.

PET is commonly used with computer tomography (CT) or magnetic resonance imaging (MRI) since PET is valuable for functional imaging; on the other hand, CT and MRI add value to anatomical reference imaging [3]. CT is also used for attenuation and scatter correction in PET/CT studies.

Typical clinical applications of FDG PET/CT are benign/malign differentiation, staging, monitoring the effect of therapy, posttreatment following, detecting tumor recurrence, selecting the region for biopsy, and guiding radiation therapy.

1.2 Non-Small Cell Lung Cancer (NSCLC)

Lung cancer is one of the most frequently diagnosed cancer types worldwide. Because of the lack of clinical symptoms and effective screening, lung cancers are diagnosed at an advanced stage. This makes lung cancer the leading cause of cancer-related death. Staging lung cancer in its initial stages is vital since treatment procedures and prognosis evaluation depend on it. The staging system is based on the Tumor-Node-Metastasis (TNM) classification form, which is announced and updated by the International Association for the Study of Lung Cancer committee based on an evaluation of the literature and clinical examination worldwide [4].

Almost 85% of lung cancers are non-small cell lung cancer (NSCLC) [5], and adenocarcinoma (ADC) and squamous cell carcinoma (SqCC) are the two major subtypes of NSCLC. ADC and SqCC correspond to about 40% and 25-30% of lung cancers, respectively [6].

PET is a valuable functional imaging method. Its efficiency for patients with cancers of NSCLC to stage tumors, evaluate therapy response, define prognosis, and guide radiotherapy and surgery is proven.

1.3 Tumor Heterogeneity

The Assessment of tumor heterogeneity is vital for the therapy. Cancer is a progressive disease; through time, cancer turns more heterogeneous. Due to heterogeneity, the sensitivity of treatment is differentiated.

Cancer transforms nonmalignant tissue into malignant by breaking the key cellular processes. The progression of cancer does not follow a linear process; instead, its nature is stochastic. Due to the dynamic nature of cancer, molecularly heterogeneous bulk tumors include different levels of cells whose reaction to treatment is different [7].

1.4 Radiomics

The central hypothesis of radiomics is the following: medical images include more information than may be obtained by visual analysis [8]. Thanks to the increase in PET scanners' spatial resolution, researchers tend to use image processing tools/approaches to PET images. In this perspective, features extracted from PET images may help us describe certain tumor properties in vivo at the molecular level.

Different textural features and automatic classification approaches have been utilized in different contexts, such as predicting response to therapy and survival [2,3] tumor grade [9]. Texture analysis (a subset of the radiomics) is an approach that includes a set of pattern recognition and analysis methods. These methods are used to quantify the relationship between the pixels or voxels for better tumor characterization, monitoring, and predicting therapy response and prognosis. Computed tomography (CT) images have also been used for pulmonary nodule feature optimization [10], reproducibility and prognosis [11], and predicting survival [12].

Although radiomics in cancer research has been a hot topic in the last decade, there are no robust features offered by the scientific community to be used instead of a PET parameter called Standardized Uptake Value (SUV) in the clinical routine [4, 6, 8–19]. This is due to the complexity of cancer biology and inter or intra variability of cancer. There are various challenges in tumor characterization using image processing approaches. To offer such a radiomics feature, tumor heterogeneity must first be analyzed.

For this purpose, the first step is noise removal on images. Each imaging modality introduces a different noise due to different image acquisition techniques [20–25]. Secondly, delineation or segmentation of the tumors in three dimensions is needed. Then, from the segmented tumor images using various methods, radiomics features are extracted. Finally, using statistical analyses and machine learning techniques features are analyzed to obtain more information about the nature of the tumors. Figure 1.3 illustrates the general framework of the cancer radiomics study.

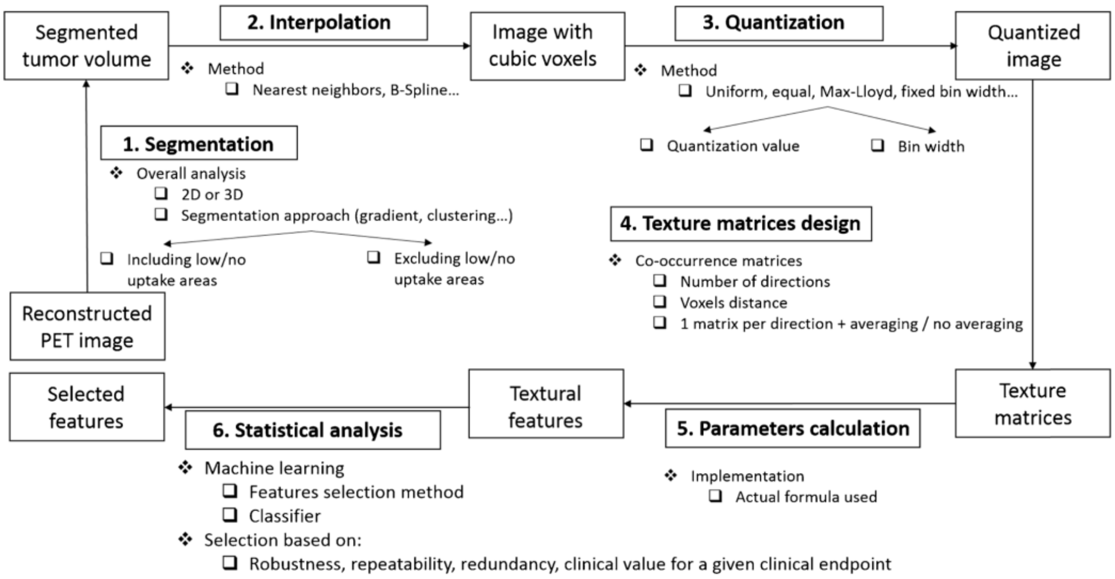


Figure 1. 3 General workflow of radiomics studies [32]

The second step of the workflow is the segmentation. Segmentation is the separation of a region of interest from the image. The idea behind the segmentation is to model the image as a function and solve the equation based on constraints. Various types of models exist in the literature [33]–[36]. However, we will focus on the random walk algorithm derived from graph cut theory because it performs well on noisy and fuzzy images, although this approach has various challenges. The user must put a seed on the image, which might introduce some level of variability. Different seed points may result in different segmentation results.

After segmenting the tumor in all slices, we create a three-dimensional matrix with the help of interpolating pixel values in the slices [32]. Later, we need to quantize the image, which corresponds to the process of representing a constant value of pixels on the image by a set of discrete values.

Finally, we will extract various features from three-dimensional segmented and quantized tumor matrix. We can categorize features based on frequency, statistical, fractals, harmonics, and probabilistic methods [37]. Although there are many opportunities for extraction of features, the curse of dimensionality makes things more challenging. Before choosing a classifier, we need to normalize data and consider the correlation between features. One feature subset selection method is needed, such as sequential forward selection. This step affects classifier performance also. That is why we will consider selecting feature and classifier together [35–42]. We will apply several classifiers and compare the result.

Each step in the framework may cause concern about stability and interpretability. However, physicians have been using SUVs since verifying image-based features is not good enough. In addition, we should note that most studies ignore features' repeatability and robustness. There are several factors of consideration.

1.5 Deep Learning

Figure 1.4 represents the conventional CNN architecture. CNN consists of several layers; each layer has a specific task. Common layers of CNN are convolution, pooling, activation, and fully connected layer. Each layer gets input from the previous layer. The convolutional layer consists of a set of filters that help extract features. Convolution is an elementwise sliding window operation of a filter applied on input. The output of each filter is called an activation map. Stacking each activation map is the output of the convolutional layer. Each filter extracts distinctive features, so each filter activates a different image region. Filter size defines a neuron's receptive field, which is vital when working with an image. If we take each pixel as an input, neuron computation complexity increases exponentially. However, we can down-sample the region using a convolutional layer while increasing the output depth. Control of the output convolution layer size has parameters such as depth, stride, and pooling. The output size of the convolutional layer of an input ($W_{input} \times H_{input} \times D_{input}$) can be obtained as;

$$W_{output} = ((W_{input} - F + 2P) / S) + 1 \quad (1.1)$$

$$H_{output} = ((H_{input} - F + 2P) / S) + 1 \quad D_{output} = K \quad (1.2)$$

The activation layer applies the chosen activation function (linear or non-linear) to the input and does not change the input size. Where F is receptive to field size, S is a stride parameter, and P is the pooling parameter. The pooling layer has two crucial contributions: first, to reduce spatial resolution, which reduces the parameters of CNN; second, to prevent overfitting. A fully connected layer is generally placed at the end of the network before the classifier.

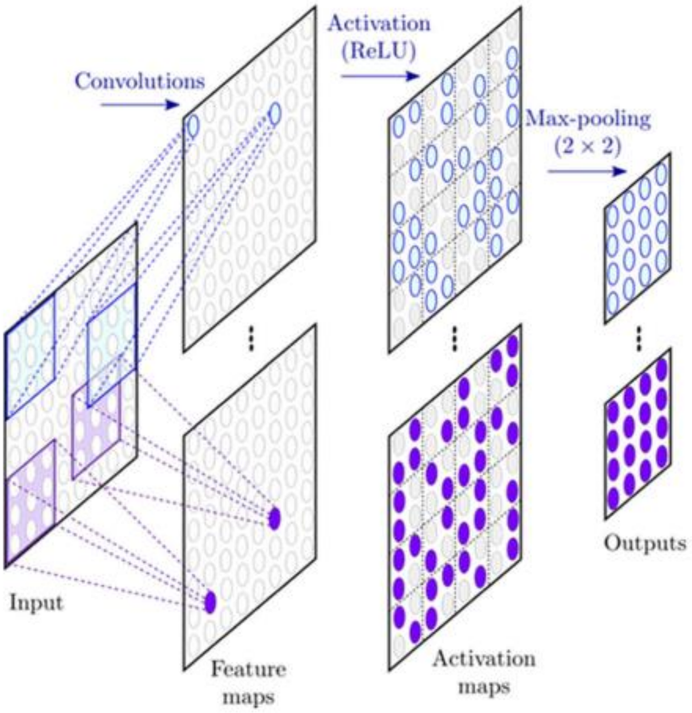


Figure 1. 4 Conventional CNN architecture [46]

CNN can be transformed to solve the segmentation problem. Long et al. [47] developed fully convolutional network (FCN) using CNN to pixel-wise prediction. For pixel mapping, They used higher resolution maps combined with up-sampled version of the previous convolution layer. The output of the FCN is the same size as the input, and each pixel is assigned to a class. The structure of FCN is given in Figure 1.5.

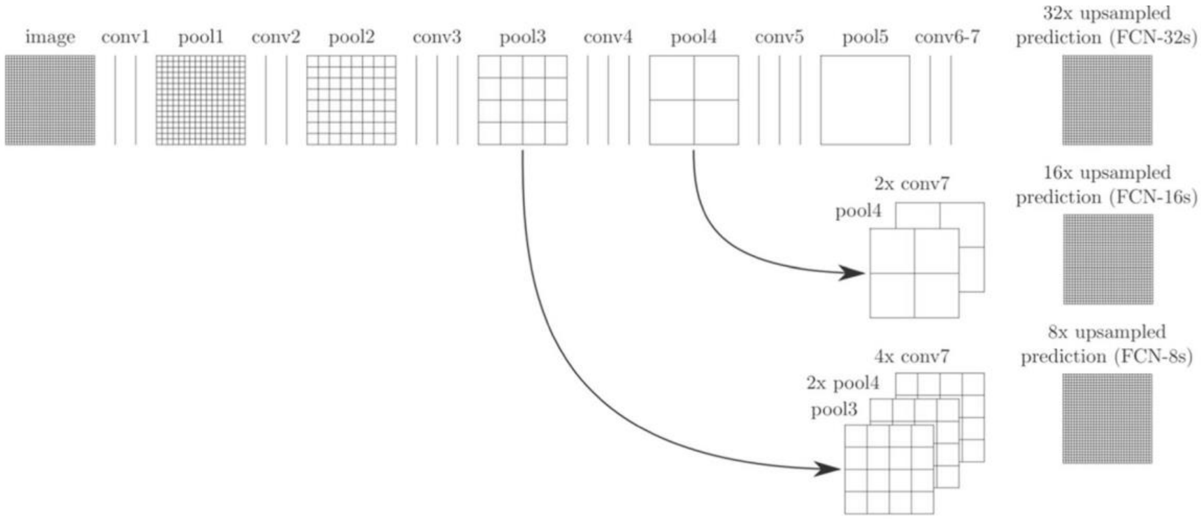


Figure 1. 5 FCN architecture [48]

Ronneberger et al. [49] developed another approach that is quite popular in medical image segmentation. U-Net architecture introduced deconvolution and skip connection approach to improve FCN architecture. In Figure 1.6, you can see the U shape structure, the first part of the architecture encodes the information, and the next part decodes the image with the help of a skip connection.

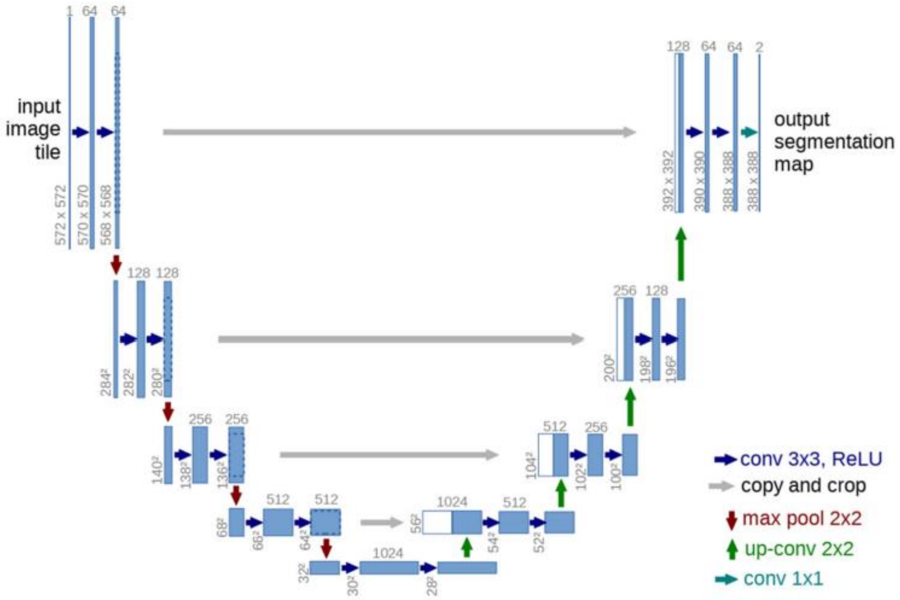


Figure 1. 6 U-Net architecture [49]

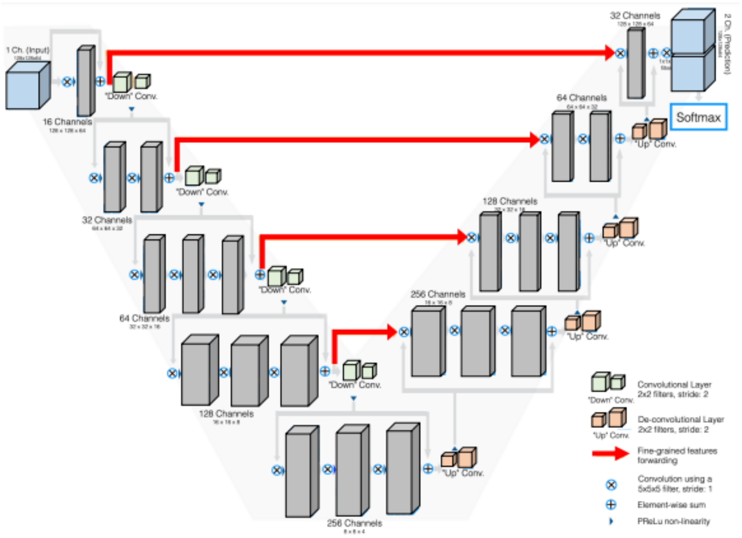


Figure 1. 7 V-Net architecture [50]

The concept of deep learning (DL) aims to create a model that transforms input data into output data using layers. Generally, each model has an input layer, hidden layers, and an output layer. The most common approach in image processing is convolutional neural networks (CNNs). CNN performed better than hand-crafted features in problems related to image and speech fields. The main layers are convolution, pooling, and fully connected layers. These layers extract higher-level features from input. At each layer, the layer's input convolves with a kernel, adds a bias parameter, and generates a new input for the next layer.

CNN is a subcategory of deep learning. Although CNN has a long history, going back to the 80s, the topic got attention with AlexNet, which won the ImageNet challenge in 2012. There are two main reasons; first, GPUs (faster than CPUs) were introduced. Secondly, open-source software platforms were created for research and commercial use. Because of commercial benefits, Google and Facebook led the artificial intelligence research and created libraries and platforms for deep learning. Medical image application of the DL can be categorized into classification, detection, segmentation, detection, and image generation. Besides the problem of definition of the interested organ, region and modality were used to create enormous literature because each problem has its own challenge. For instance, PET images have low resolution; on the other hand, CT images have higher resolution, and brain tumors and lung tumors have different challenges.

Although DL approach has a tremendous effect on natural image processing, medical image processing has its own challenges. DL became successful thanks to a substantial number of training samples; however, medical field has the problem of limited annotated data. Most datasets are not available for public use, and it is hard to find a clinician to volunteer to annotate. Even if the dataset is available, different imaging modalities and diseases create small data sets compared to natural images. One of the viable solutions is data augmentation which increases the performance by increasing the number of samples with random transformations. In medical images, augmentation helps to improve performance in a limited way compared to natural images. Another solution is to apply transfer learning, which learns the architecture parameter in a different dataset and then fine-tune the parameter based on a small dataset. Even natural and medical images are different; filters and layers can capture key features. Another challenge is the class imbalance problem; for instance, in the tumor detection problem and tumor size is considered insignificant compared to the background tissue. This fact causes certain amount of bias in the model to classify any pixel as background and still get high accuracy values. One can use different loss functions to overcome the class imbalance in segmentation problems, in which weights of class matter.

Medical image segmentation is a step that extracts the region of interest based on problem definitions such as organ, tumor, and lesion detection. It is a vital step for dosimetry, therapy planning, and therapy response in oncological PET imaging. Automating segmentation is important because manual segmentation is a time-consuming and subjective task for the clinician.

PET/CT modality can be used in multimodality segmentation. PET and CT encode different information, so we can extract features to fuse each modality of information. However, these two modalities have different resolutions and spatial size and require image registration to match the exact same location. It is not a straightforward process since medical images are not suitable for rigid transformation; most of the body parts are deformable and do not have exact geometric shapes. Problem definition is important for segmentation. For instance, the grade classification multimodality image segmentation effect can be observed. Creating an experiment set-up shows each modality's statistical power to classify the grade of the tumor. However, developing a different model and fair comparison is difficult to develop each modality. In radiomics, studies show that segmentation is the first step to extracting features.

The formulation of radiomics is well defined, and applying the same formulation for different modalities is hard to interpret. For example, the contrast features of PET and CT images are different based on classification technique, one of the same feature attribute performances. In radiomics performance, evaluation is affected by segmentation techniques, pre-processing, feature selection approach, and classification techniques. Each step has its own drawback. It makes it hard to interpret the result, which is significant in medical research. If we cannot explain to doctors, how can doctors explain to the patient how and why it works? In the end, we evaluate the pipeline based on performance. Translating research to clinics is a crucial purpose. Society needs more annotated data, multi-center cancer images, and medical image challenges to transfer models to clinics. In our study, we focused on only PET segmentation and primer tumor. There are two reasons: first, cancer is a complex disease, that is why we have to isolate cases to explain it clearly, and second, metabolic activity evaluation is critical for oncology.

There are many review papers for segmentation in medical images for deep learning applications [46], [51]–[53]. Each review focuses on the different advantages of techniques. One common point, the encoder-decoder approach, is preferred over LSTMs, GANs, and RNNs. There could be several reasons, such as DL research evolves different directions and translating computer vison DL research to medical images takes time.

Another reason is that limited annotated data make them hard to compare fairly. Lastly, encoder-decoder architectures gave satisfactory results, and there is no way to compare all possible architectures in the same dataset. Among the encoder-decoder family, the most used architectures are U-Net, V-Net, and 3D U-Net. The general approach to the problem is several decision steps. First, the researcher must decide to work on volumetric or slice-based data.

For our case, we chose a slice-based approach that could be better. Because we already have one nuclear medicine doctor and we hypothesized manual segmentation is error-prone. If we use volumetric data, subjective segmentation affects the result more than the slice-based data. Another reason is that the volume of the tumor varies in our dataset. To apply the 3D approach, we need to interpolate the size of each patient. This is required for the training phase, and each input size should match the other. To clarify this, one patient has thirty slices which include a tumor. On the other hand, another patient has only six slices. Matching size based on the most significant volume creates an unrealistic volume. Then the problem is how to choose the slice for each patient. We chose slice-based of SUVMax value, which is clinically meaningful. The 2D approach can be extended for 3D and even model work, and each slice can be segmented and then could be stacked to obtain volumetric data, called the 2.5D segmentation approach.

After deciding on the slice-based approach, the next step is architecture family. As mentioned earlier, encoder-decoder architectures are popular in the medical field. In natural image segmentation, according to the technical contribution, model families can be classified up to ten. Among them, we choose dilated convolutional models and DeepLab family. The main reason is that DeepLabv3 has achieved an 89% mIoU score on the PASCAL VOC challenge.

1.6 Purpose of the Thesis

SUV is a semi-quantitative image biomarker of tumor heterogeneity in FDG PET/CT imaging system. Repeatability and reproducibility of image biomarkers are essential for patients' clinical management. Repeatability is yielding the same result in the same patient being examined on the same system. In contrast, reproducibility is the ability to produce the same result in the same patient being examined on a different system. SUV has a problem in terms of standardization since it is affected by reconstruction method, scanner parameters, biological factors, and the imaging equipment [54]. The European Association of Nuclear Medicine (EANM) published guidelines to standardize PET tumor imaging to help physicians standardize diagnostic quality and quantitative information in oncology patients [55]. However, the final judgment of the procedure is made by medical professionals, and this situation prevents SUVs from being robust biomarkers and varying from center to center. Also, it creates contradictory heterogeneity, resulting in literate because of reproducibility.

Clinicians need more robust features. Our first goal is to extract features from PET images to replace or can be used with SUVs. Our second purpose is to develop and test machine learning approaches to classify lung cancer subtypes. It is important since the golden standard of diagnosing subtypes of lung cancer based on biopsy is an invasive technique, and future targeted therapy will be based on subtype-specific.

This study investigates tumor heterogeneity with image processing and pattern recognition techniques to develop robust features that can be used in clinical routine. In the first part of this thesis, radiomics properties, machine learning, and feature selection were investigated in the subtype determination of lung cancer. In the second part, the tumor segmentation in the expanded patient dataset was examined, and a model with a deep learning approach was applied, achieving success close to manual and semi-atomic methods. In these two parts, we used images from 154 patients with NSCLC that previously underwent 18F-FDG-PET/ CT imaging for cancer staging before surgery, chemotherapy, or radiotherapy treatment according to the stage of their disease from March 2010 to April 2014 evaluated in Acıbadem Kayseri Hospital. Patients were grouped as stage I, II, III, or IV, using conventional CT criteria for tumor size and local invasion and PET assessments of nodal and distant metastases by well-trained imaging specialists according to the seventh edition of the American Joint Committee on Cancer (AJCC) TNM classification guidelines. Malignant disease was confirmed by histopathological verification in all patients. In the last part, PET images from 72 patients with pancreatic cancer were analyzed to determine whether the tumor characteristics effectively determine the life span.

Chapter 2

Adeno and Squamous Cell Lung Cancer Differentiation

2.1 Background Information

Until recently, therapeutic approaches to NSCLC were guided by tumor stage, and there was no difference in treatment for ADC vs. SqCC. The significant advances in understanding the effects of cytotoxic and biological agents used in the NSCLC therapy suggest that future targeted therapies will be increasingly subtype-specific. Selection of patients for appropriate subtype-specific therapies requires precise pathologic differentiation of ADC and SqCC [56]. The lung cancer diagnosis is usually performed based on small biopsy (bronchoscopic, needle, or core biopsies) and cytology specimens. Usually, these two subtypes are distinguished based on standard morphologic criteria by routine microscopy. However, distinguishing can be difficult in some poorly differentiated tumors, especially small specimens. On the other hand, the characterization of the lesion using a small biopsy might have a sampling error, which would not represent the actual biological behavior and the intratumoral heterogeneity.

Positron emission tomography (PET) is a valuable functional imaging method. Its efficiency for patients with cancers of NSCLC to stage tumors, evaluate therapy response, define prognosis, and guide radiotherapy and surgery is proven. Recently, a concept called radiomics has become popular. The central hypothesis of radiomics is that medical images include more information than may be obtained by visual analysis [57]. Thanks to the increase in PET scanners' spatial resolution, researchers use image processing tools/approaches to PET images. In this perspective, features extracted from PET images may help us describe certain tumor properties in vivo at the molecular level. Texture analysis is an approach that includes a set of pattern recognition and analysis methods. These methods quantify the relationship between the pixels or voxels for better tumor characterization, monitoring, and predicting therapy response and prognosis. Different textural features and automatic classification approaches have been utilized in different contexts, such as predicting response to therapy and survival [2, 3] and tumor grade [17]. Computed tomography (CT) images have also been used for pulmonary nodule feature optimization [10], reproducibility and prognosis [8], and predicting survival [60]. In addition to medical imaging approaches like PET and CT, for lung cancer diagnoses, automated quantitative analysis of histopathology images has been investigated [48, 49].

Machine learning studies the construction of algorithms that can learn from and make predictions on data to make intelligent decisions based on their recognition of complex patterns. Machine learning methods are used in oncology in different applications such as cancer prognosis and prediction [63], survival analysis [38], drug response [64], and gene expression [65]. The focus of this study is medical image analysis and computer-aided diagnosis. This classification problem uses PET images to determine whether a newly presented patient has a tumor subtype adenocarcinoma or squamous cell carcinoma. Thus, the oncological therapy may be guided accordingly. In a similar study [66] that aimed to cluster the subtypes using 24 textural features obtained from the PET images, the researchers used linear discriminant analysis as the classification approach.

In this study, we have used 39 textural features frequently preferred by researchers to characterize the tumor heterogeneity and analyzed the performances of different classification approaches that have not been utilized in the tumor subtype discrimination in NSCLC.

2.2 Materials and Methods

2.2.1 Patient population and PET/CT imaging

This study includes 18F FDG PET/CT images of 96 patients with non-small cell lung cancer (NSCLC). The imaging of patients was performed from March 2010 to April 2014 at Acıbadem Kayseri Hospital Nuclear Medicine Department, Kayseri, Turkey, using a PET/CT scanner (Siemens Biograph 6, HiRez). The Research Ethics Committee of Kayseri Research and Training Hospital (KRTH) approved this study. Out of 96 patients, 8 were females, and 78 were males, with a mean age of 62.9±4.5 (range: 39-84). The tumor subtypes of 36 patients were ADCs, and 60 patients were SqCCs. The specimens were obtained using fine-needle or excisional biopsy and were assessed at the pathology department of KRTH in terms of tumor subtype.

2.2.2 Image processing and texture analysis

For each patient, PET and CT images were transferred to our computers. This study was focused on the PET images, especially slices with tumors. The MATLAB (MathWorks MA, USA) program was used in the image processing steps of PET images in DICOM format. In the image processing part of the study, first, the tumors were segmented in each slice, the image intensity values in the tumors in that slice were binned, and finally, texture analysis approaches were applied to extract texture features from each three-dimensional tumor obtained by arranging two-dimensional slices in one stack. In the segmentation, a popular approach called random walk (RW) [36] was used to distinguish the tumor from the background automatically. Different segmentation methods like Otsu's, k-means, active-contour approaches were also tested, and the best results were obtained using the RW approach. The binning process corresponds to a linear mapping of intensity values on the pixels of the segmented tumor region to be between 1 and 64. Various binning levels were tested, and 64 was found to be the optimal value, as [16] proved that levels more than 64 do not improve classification precision. In the last step, four different texture analysis approaches from the binned regions with tumors' 39 features were extracted. The approaches we used were the gray level co-occurrence matrix (GLCM, 8 features), gray level run length matrix (GLRLM, 13 features), gray level size zone matrix (GSZM, 13 features), and neighborhood gray-tone difference matrix (NGTDM, 5 features). The details of these approaches can be found in [67]. The most common quantitative value derived from PET images that shows radiotracer uptake is the maximum standardized uptake value (SUVmax) in the tumor area, defined as the decay-corrected tumor activity concentration divided by injected activity per unit body weight, surface area, or lean body mass. In addition to the textural features, we have also included the SUVmax as the 40th feature, whose values ranged from 2.5 to 47.1 (15.5±7.4).

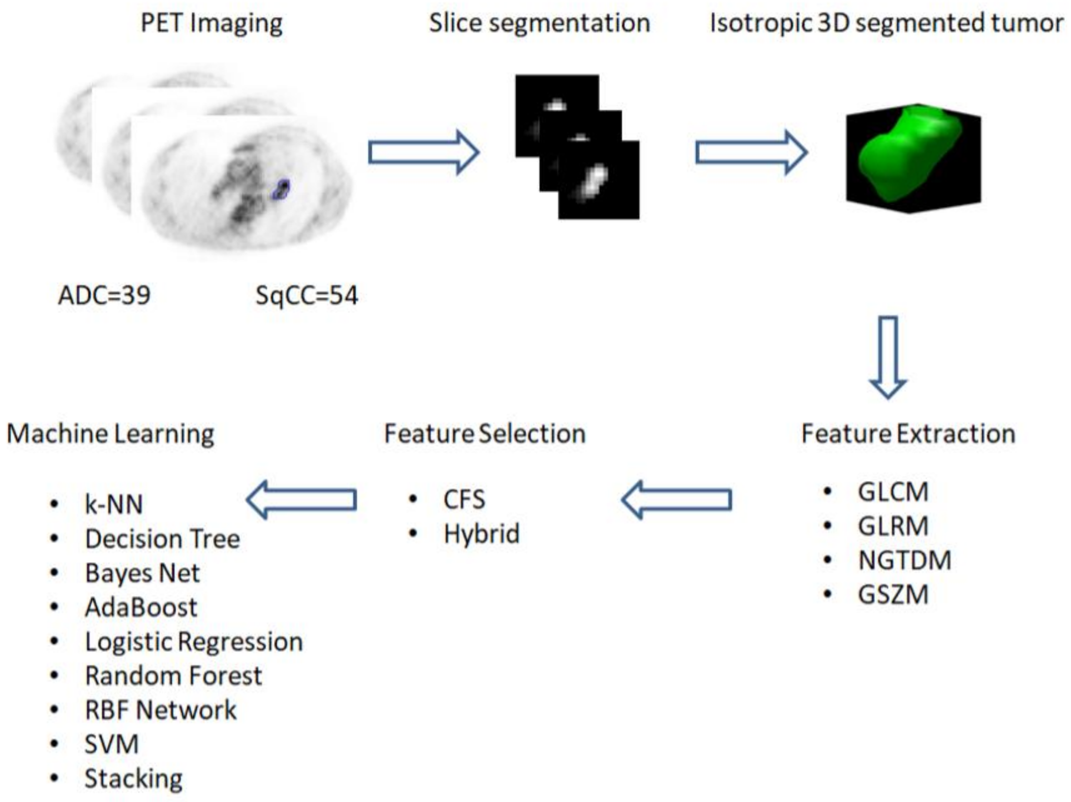


Figure 2.1 Summary of the approaches used in this study

2.2.3 Data preprocessing and feature selection

We considered normalizing the texture features to the interval from 0 to 1. Feature selection methods are classified into filter, wrapper, and ensemble feature selection. The influence of the feature selection method on the performance of the classification method was examined before in [65], and it was found that ensemble feature selection does not improve accuracy generally on breast cancer prognosis. To reduce the number of dimensions, we implemented two feature selection methods in WEKA [68]: (1) CFS subset evaluator with BestFirst search strategy and (2) a hybrid strategy that first ranks features according to gain ratio and followed by a wrapper method that selects features using the k-NN classifier (with k parameter optimized by 10-fold cross-validation).

PET system introduces Poisson noise to the resultant image due to the stochastic nature of the photon counting process at the detectors, and this noise is signal-dependent. For instance, thermal and electronic fluxions of the acquisition system are signal independent, and one can assume that additive Gaussian noise. Most relevant work either has not implemented any preprocessing approaches or assumed that noise is additive Gaussian noise. One of the common methods for preprocessing applied on PET images is the “Anscombe’s method,” which is based on variance stabilization. In this approach, signal-dependent Poisson noise can be modeled as independent additive Gaussian noise. Studies show that preprocessing increases SNR [26-29] and PSNR [69]. No recent work indicates this preprocessing approach's effect in this area. We asked an expert about our dataset, and the SNR level is enough to proceed. We did not apply the image denoising algorithm to our dataset.

2.2.4 Classification methods

In the present study, we implemented 11 different classifiers in WEKA software to differentiate the ADC and SqCC tumor subtypes: k-nearest-neighbor (k-NN), logistic regression, support vector machines (SVM), Bayesian network, decision tree, radial basis function (RBF) network, random forest, AdaBoostM1, and three stacking methods. We chose these classifiers for the same reason as Parmar et al. [38], due to their popularity in the literature. We performed a leave-one-out cross-validation (LOOCV) experiment on the dataset to evaluate the prediction accuracy. We also considered optimizing certain hyper-parameters of these models by performing 10-fold cross-validation separately on each training set.

2.2.4.1 k-nearest neighbor

k-nearest neighbor (k-NN) classifiers first find the k training samples closest to the test example and combine the class labels of these nearest neighbors by majority voting [22]. We employed the IBk method in WEKA to implement the k-NN classifier in our experiments. We considered selecting the number of nearest neighbors (i.e., the k parameter) as 3 as well as optimizing this parameter by including the –X option in the command line choosing the maximum number of nearest neighbors as $number\ of\ samples - 2$ and setting the number of cross-validation folds to 10.

2.2.4.2 Logistic regression

As a special case of generalized linear models, the logistic regression classifier computes a weighted linear combination of input features passed through a non-linear activation function (e.g., a sigmoid). The class labels are assigned in binary classification by comparing the output variable to 0.5. The decision boundaries of a logistic regressor are linear hyperplanes [41]. We employed the logistic classifier in WEKA, which implements a multi-nomial logistic regression method with a ridge estimator and Quasi-Newton optimization procedure.

2.2.4.3 Support vector machines

A support vector machines (SVMs) classifier aims to solve a quadratic optimization problem [40] by mapping the training samples to a higher dimensional space and finding a linear separating hyperplane with a maximum margin [45]. We implemented two SVMs with a Radial Basis Function (RBF) kernel using the LIBSVM package [24] in WEKA. In the first version, we set the C parameter to 1.0 and γ to 1/number of features, while in the second model, we optimized these hyper-parameters by performing a grid search, choosing $C \in (2^{-5}, 2^{-4}, \dots, 2^{15})$ and $\gamma \in (2^{3}, 2^{2}, \dots, 2^{-15})$. At the end of this procedure, we selected the particular pair that gives the best cross-validation accuracy, trained the SVM classifier using these optima and performed predictions on the test sample.

2.2.4.4 Decision tree

A decision tree classifier contains nodes and directed edges (i.e., branches) connecting nodes with no cycles allowed. Each internal node represents a test on a feature, and each branch is the outcome of the test, which can be true or false. For a given feature vector, the tests are applied from the top (root) node down to the leaf nodes, representing a class label (i.e. final decision). Hence, each path from the root to a leaf node is a classification rule. We employed the J48 algorithm in WEKA (a successor of C4.5) under default parameters, in which the confidence threshold for pruning is set to 0.25, and the minimum number of instances per leaf is set to 2 [70].

2.2.4.5 Bayesian network

Let $X = [x_0, x_1, x_2, \dots, x_d]$ be the set of variables, where $x_0 = y$ is the output class variable and $x_1, x_2, \dots, x_d$ represent input features. A Bayesian network B over variables in X is a directed acyclic graph (DAG) and a set of probability tables $B_P = \{p(x|pa(x))|x \in X\}$ where pa(x) is the set of parents of x. The probability distribution for X can be computed as $P(X) = \prod_{x \in X} p(x|pa(x))$. The classification problem can be stated as inferring the class variable $y = x_0$ given the set of input features $\mathbf{x} = [x_1, x_2, \dots, x_d]$. In this context, a BayesNet classifier $f : \mathbf{x} \rightarrow y$ is a function that maps an input feature vector x to class type y. The classifier is learned from a dataset containing samples over (x, y), and the learning process includes deriving a Bayesian network structure and the mapping function f. The classification process selects the particular class type that maximizes the a posteriori distribution $P(y \mid \mathbf{x})$. In the present study, we employed the BayesNet classifier in WEKA software, which first discretizes the continuous-valued features by employing the filter called weka.filters.unsupervised.attribute.NumericToNominal. We selected the search algorithm for learning the network structure as K2, a hill-climbing algorithm restricted by the order of the variables, and the estimator as SimpleEstimator, which computes the conditional probability tables (CPTs) directly from the data for a given network structure [43].

2.2.4.6 Radial basis function (RBF) network

A radial basis function network first clusters data and then fits a basis function to each cluster. In the second stage, the basis function outputs are sent to a linear classifier to predict the class type [71]. We employed the RBFNetwork classifier in WEKA, which uses the k-means clustering algorithm and fits symmetric multi-variate Gaussians to data in each cluster. The output of Gaussians, which constitute the basic functions, are directed to a logistic regression classifier to predict the class type. All data are normalized to zero mean and unit variance (i.e., Z-score normalization). We implemented two versions of the RBFNetwork. The first one uses two clusters, which are equal to the number of class types, and the second optimizes the number of clusters by cross-validation considering the following values: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 25.

2.2.4.7 Random Forest

A random forest classifier is an ensemble technique that combines multiple decision trees by weighted majority voting. Each tree receives a small subset of input features constituted by random selection and is trained on a separate training set, which is generated by the bootstrap sampling procedure (also known as bagging) [42]. Random forest is also robust against outliers and is less prone to overfitting. We implemented two versions of the RandomForest classifier in WEKA. The first one uses 100 trees, and the second one optimizes the number of trees by performing cross-validation on each training set and considering the following alternatives for this parameter: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 25, 30, 35, 40, 45, 50, 60, 70, 80, 90, 100.

2.2.4.8 AdaBoost

A boosting ensemble combines multiple classifiers through weighted averaging of classifier outputs. Different from bagging, the base learner at a given iteration is constructed according to the classification behavior of the previous learner concentrating more on the misclassified examples. To construct the training set of the current classifier, the probability of selecting misclassified examples is increased, and a bootstrap sampling procedure is used [44]. Although boosting can be prone to overfitting, it typically improves the overall classification accuracy. We employed the AdaBoostM1 method in WEKA by selecting DecisionStump as the base learner and implemented two versions of this classifier. The first one selects the number of iterations as 10, which is the default value. The second optimizes this parameter by performing cross-validation on each training set considering the following values: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 25, 30, 35, 40, 45, 50, 60, 70, 80, 90, 100.

2.2.4.9 Stacking

A stacking ensemble combines different types of classifiers, which serve as base learners through a meta-learner [72]. Typically, the number of base learners is smaller than bagging or boosting. We implemented three stacking ensembles in the present study by combining different classifiers. The first ensemble combines the decision tree (i.e., J48 in WEKA) with AdaBoostM1 (Stacking 1), the second combines the decision tree, AdaboostM1, and logistic regression (Stacking 2), and the third combines the decision tree, AdaboostM1, logistic regression and BayesNet classifiers (Stacking 3). We employed logistic regression as the meta-learner in each method and used 10 iterations for AdaBoostM1, the default setting in WEKA.

2.2.4.10 Performance Measures

We used the following measures to evaluate the performance of the classifiers: Sensitivity (or recall), specificity, positive predictive value (PPV), negative predictive value (NPV), Matthew’s correlation coefficient (MCC), F-measure, overall accuracy, and area under ROC curve (AUC) [21]. These are computed as

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2.1}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{FP} + \text{TN}} \tag{2.2}$$

$$\text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2.3}$$

$$\text{NPV} = \frac{\text{TN}}{\text{TN} + \text{FN}} \tag{2.4}$$

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}} \tag{2.5}$$

$$\text{F} - \text{measure} = \frac{2 \times \text{precision} \times \text{recall}}{(\text{precision} + \text{recall})} \tag{2.6}$$

$$\text{Overall Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{2.7}$$

where TP is true positives, FP is false positives, TN is true negatives, and FN is false negatives. AUC measure is computed by first ranking the predictions with respect to the decision scores and then shifting the decision threshold to compute TP, FP rate values of the ROC curve. Each horizontal move (i.e., a false positive) generates a rectangular region in ROC, and the cumulative sum of these areas gives our AUC estimate.

2.3 Results and discussion

We performed a leave-one-out cross-validation experiment on the main dataset and obtained the accuracy measures shown in Tables 2.1 to 2.3. Table 2.1 compares different classifiers when no normalization is applied, and the hybrid feature selection strategy is used. Table 2.2 demonstrates the accuracy of classifiers when data is normalized, and hybrid feature selection method is employed. Table 2.3 includes the accuracy measures of the stacking ensemble for all combinations of the following conditions: Data is not normalized, data is normalized, no feature selection is performed, CFS subset evaluator is employed, and a hybrid feature selection method is employed.

According to these results, we achieved the best results with the decision tree approach and stacking classifiers when data is normalized, and the hybrid feature selection is used. Because the decision tree was among the base learners in all stacking methods implemented, we can conclude that stacking ensemble does not improve the accuracy of its base learners further. Based on the results presented in Tables 2.1 to 2.3, we can also observe that feature selection generally increases the classification accuracy compared to the condition where no feature selection is employed. Comparing the two feature selection methods, some classifiers are more accurate when the first feature selection method is used, while the rest gives better results with the second strategy. Similar behavior is observed for data normalization conditions, and no winner takes all conditions that exist. Furthermore, hyper-parameter optimization improved the prediction accuracy of certain classifiers but not all of them. This could be related to constraints imposed by having a small number of samples. Table 2.4 shows the confusion matrix for the LOOCV experiment. It is evident that when the tumor subtype is SqCC, the prediction is more successful, but the identification is harder for the ADC subtype.

Figure 1 shows the histogram of the number of features selected on each training set of the leave-one-out cross-validation (a total of 96 feature subsets) when the hybrid feature selection is employed, and no normalization is applied. According to this figure, most of the time, approximately 20 features are selected out of 40. Similar behavior is observed when the same experiment is repeated on normalized data.

Figures 2.3 and 2.4 show the relative importance of the features when the hybrid feature selection method is employed on not normalized and normalized data, respectively. The horizontal axis shows the features used in this study and the vertical axis represents the number of times a feature is selected when feature selection is repeatedly applied on each training set of the leave-one-out cross-validation. Comparing these plots, the key features are similar for the two normalization conditions.

Finally, when the decision tree classifier is trained on the normalized version of the dataset with 96 samples (without performing any feature selection), the tree diagram shown in Figure 4 is obtained, which performs a test on a single attribute named RLV (run-length variance, a parameter extracted from gray-level run-length matrix). Since a decision tree classifier inherently performs feature selection and is pruned during training, the resulting model is a feature-selected version of the original data. Furthermore, its simplicity makes it interpretable and can be applied directly in clinical settings on future data. This is also consistent with the relative importance rankings of the features in Figures 2.3 and 2.4.

Table 2.1 Accuracy measures of classifiers when no normalization is applied, and the hybrid feature selection method is employed.

Method	Sensitivity	Specificity	PPV	NPV	MCC	F-Measure	Overall	AUC
k-NN (*k*=3)	73.33	61.11	75.86	57.89	0.34	74.58	68.75	63.94
k-NN (*k* opt)	80.00	25.00	64.00	42.86	0.06	71.11	59.38	58.61
Decision Tree (J48)	66.67	30.56	61.54	35.48	-0.03	64.00	53.12	46.39
Bayes Net	83.33	36.11	68.49	56.52	0.22	75.19	65.62	53.70
AdaBoostM1 (iterations=10)	93.33	41.67	72.73	78.95	0.43	81.75	73.96	53.98
AdaBoostM1 (#iterations opt)	93.33	36.11	70.89	76.47	0.37	80.58	71.88	58.10
Logistic Regression	75.00	47.22	70.31	53.12	0.23	72.58	64.58	65.37
Random Forest (#trees=100)	73.33	38.89	66.67	46.67	0.13	69.84	60.42	59.95
Random Forest (#trees opt)	66.67	47.22	67.80	45.95	0.14	67.23	59.38	58.47
RBF Network (#clusters=15)	75.00	33.33	65.22	44.44	0.09	69.77	59.38	53.33
RBF Network (#clusters opt)	66.67	44.44	66.67	44.44	0.11	66.67	58.33	52.08
SVM default	100.00	11.11	65.22	100.00	0.27	78.95	66.67	59.86
SVM opt	73.33	36.11	65.67	44.83	0.10	69.29	59.38	55.97
Stacking 1	90.00	33.33	69.23	66.67	0.29	78.26	68.75	50.83
Stacking 2	90.00	30.56	68.35	64.71	0.26	77.70	67.71	54.21
Stacking 3	86.67	33.33	68.42	60.00	0.24	76.47	66.67	49.77

Table 2.2 Accuracy measures of classifiers when data is normalized, and the hybrid feature selection method is employed.

Method	Sensitivity	Specificity	PPV	NPV	MCC	F-Measure	Overall	AUC
k-NN (*k*=3)	68.33	47.22	68.33	47.22	0.16	68.33	60.42	56.85
k-NN (*k* opt)	86.67	22.22	65.00	50.00	0.12	74.29	62.50	55.74
Decision Tree (J48)	95.00	44.44	74.03	84.21	0.48	83.21	76.04	42.22
Bayes Net	88.33	38.89	70.67	66.67	0.32	78.52	69.79	52.31
AdaBoostM1 (#iterations=10)	91.67	38.89	71.43	73.68	0.37	80.29	71.88	49.49
AdaBoostM1 (#iterations opt)	90.00	38.89	71.05	70.00	0.34	79.41	70.83	50.79

Logistic Regression	76.67	52.78	73.02	57.58	0.30	74.80	67.71	67.31
Random Forest (#trees=100)	83.33	38.89	69.44	58.33	0.25	75.76	66.67	59.95
Random Forest (#trees opt)	71.67	58.33	74.14	55.26	0.30	72.88	66.67	61.20
RBF Network (#clusters=15)	78.33	36.11	67.14	50.00	0.16	72.31	62.50	68.75
RBF Network (#clusters opt)	53.33	47.22	62.75	37.78	0.01	57.66	51.04	56.94
SVM default	100.00	0.00	62.50	0.00	0.00	76.92	62.50	53.52
SVM opt	78.33	36.11	67.14	50.00	0.16	72.31	62.50	59.26
Stacking 1	95.00	44.44	74.03	84.21	0.48	83.21	76.04	67.18
Stacking 2	95.00	44.44	74.03	84.21	0.48	83.21	76.04	68.94
Stacking 3	95.00	44.44	74.03	84.21	0.48	83.21	76.04	62.27

Table 2.3 Accuracy of stacking methods with respect to normalization and feature selection.

Method	Sensitivity	Specificity	PPV	NPV	MCC	F-Measure	Overall	AUC
S1 FS0 N0	90.00	30.56	68.35	64.71	0.26	77.70	67.71	48.80
S1 FS1 N0	88.33	36.11	69.74	65.00	0.29	77.94	68.75	69.49
S1 FS2 N0	90.00	33.33	69.23	66.67	0.29	78.26	68.75	50.83
S1 FS0 N1	95.00	36.11	71.25	81.25	0.40	81.43	72.92	61.20
S1 FS1 N1	95.00	44.44	74.03	84.21	0.48	83.21	76.04	65.23
S1 FS2 N1	95.00	44.44	74.03	84.21	0.48	83.21	76.04	67.18
S2 FS0 N0	83.33	27.78	74.03	50.00	0.13	73.53	62.50	49.26
S2 FS1 N0	88.33	36.11	69.74	65.00	0.29	77.94	68.75	66.11
S2 FS2 N0	90.00	30.56	68.35	64.71	0.26	77.70	67.71	54.21
S2 FS0 N1	95.00	33.33	70.37	80.00	0.38	80.85	71.88	58.01
S2 FS1 N1	93.33	44.44	73.68	80.00	0.45	82.35	75.00	69.63
S2 FS2 N1	95.00	44.44	74.03	84.21	0.48	83.21	76.04	68.94
S3 FS0 N0	83.33	27.78	65.79	50.00	0.13	73.53	62.50	43.61
S3 FS1 N0	86.67	36.11	69.33	61.90	0.27	77.04	67.71	61.85
S3 FS2 N0	86.67	33.33	68.42	60.00	0.24	76.47	66.67	49.77
S3 FS0 N1	91.67	33.33	69.62	70.59	0.32	79.14	69.79	51.81
S3 FS1 N1	93.33	33.33	70.00	75.00	0.35	80.00	70.83	65.65
S3 FS2 N1	95.00	44.44	74.03	84.21	0.48	83.21	76.04	62.27

S1: First stacking method, S2: Second stacking method, S3: Third stacking method, FS0: No feature selection is performed, FS1: CFS subset evaluator is employed, FS2: hybrid feature selection is employed, N0: No data normalization, N1: Features are normalized.

Table 2.4 Confusion matrix for Stacking 2 classifier when data is normalized and the hybrid feature selection is employed.

True \ Pred	Pred = ADC	Pred = SqCC
True = ADC	16	20
True = SqCC	3	57

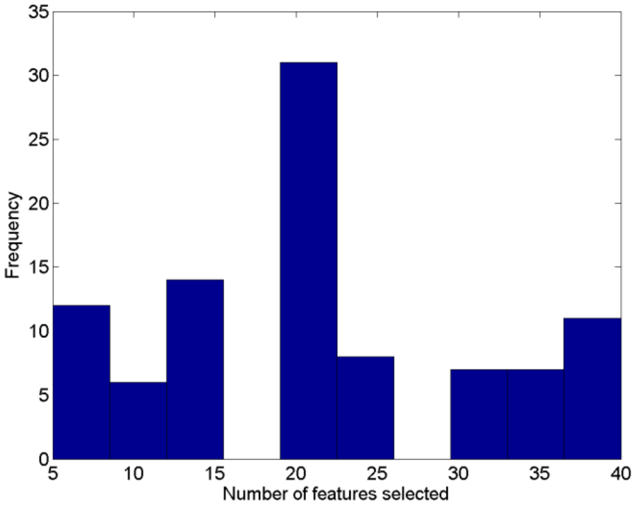


Figure 2.2 Histogram of the number of features selected on each training set of the leave-one-out cross-validation when no data normalization is performed.

In this work, we compared the accuracy of several machine learning approaches for discriminating the two cancer subtypes: adeno and squamous cell lung cancer. We also analyzed the effect of feature selection and data normalization. The most accurate method was the stacking ensemble classifier, which combines a decision tree, AdaBoostM1, and Logistic regression methods by a meta-learner. In future work, we plan to evaluate other feature selection methods in the machine learning literature and enlarge our dataset by including more subjects and new features. All these efforts are expected to advance the detection of cancer subtypes, which is very important for future targeted therapies. In addition, in the literature, this kind of discrimination problem has not been managed in such a rigorous manner from the feature selection to classification.

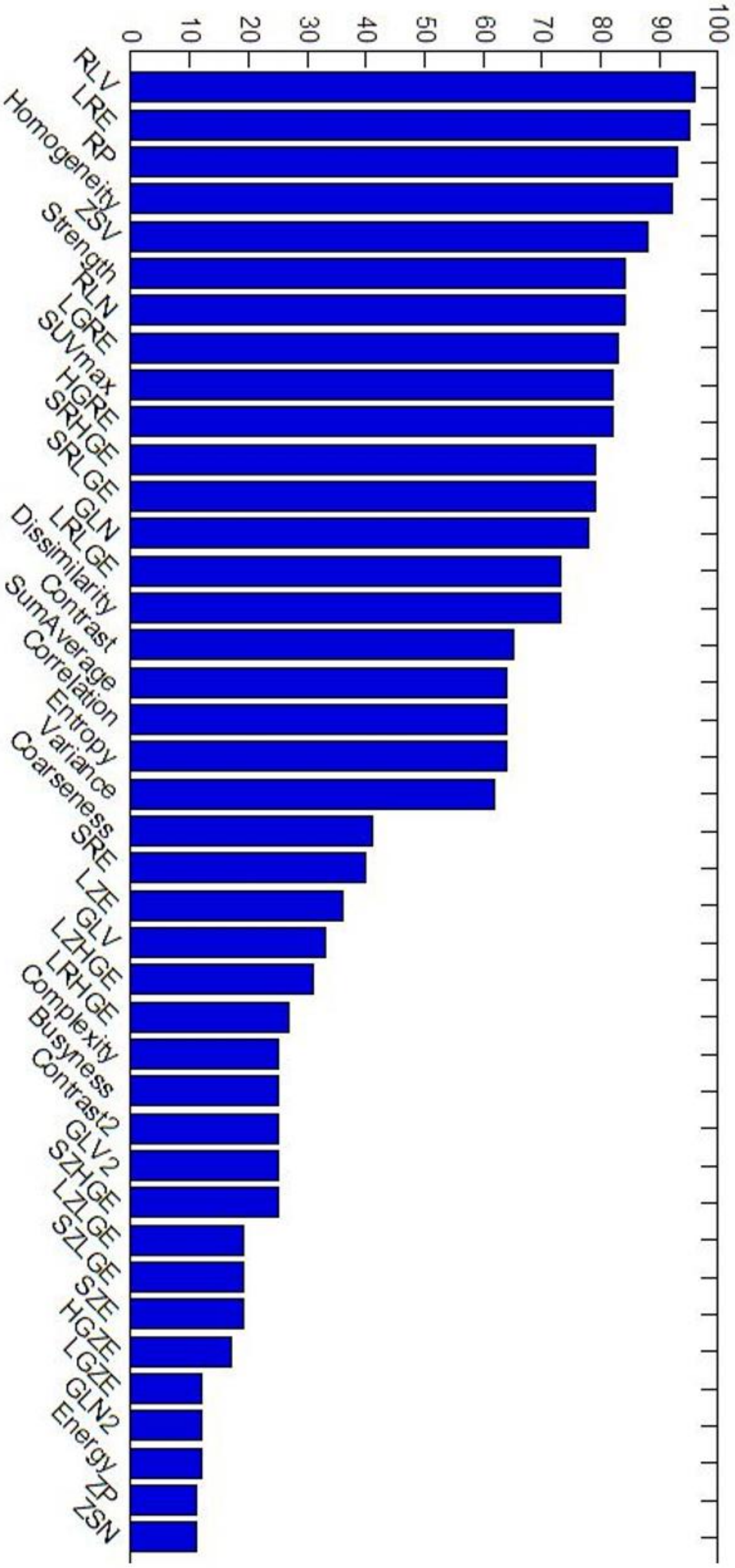


Figure 2.3 Selection frequencies of the features on training sets of the leave-one-out cross-validation when hybrid feature selection is employed, and data are not normalized.

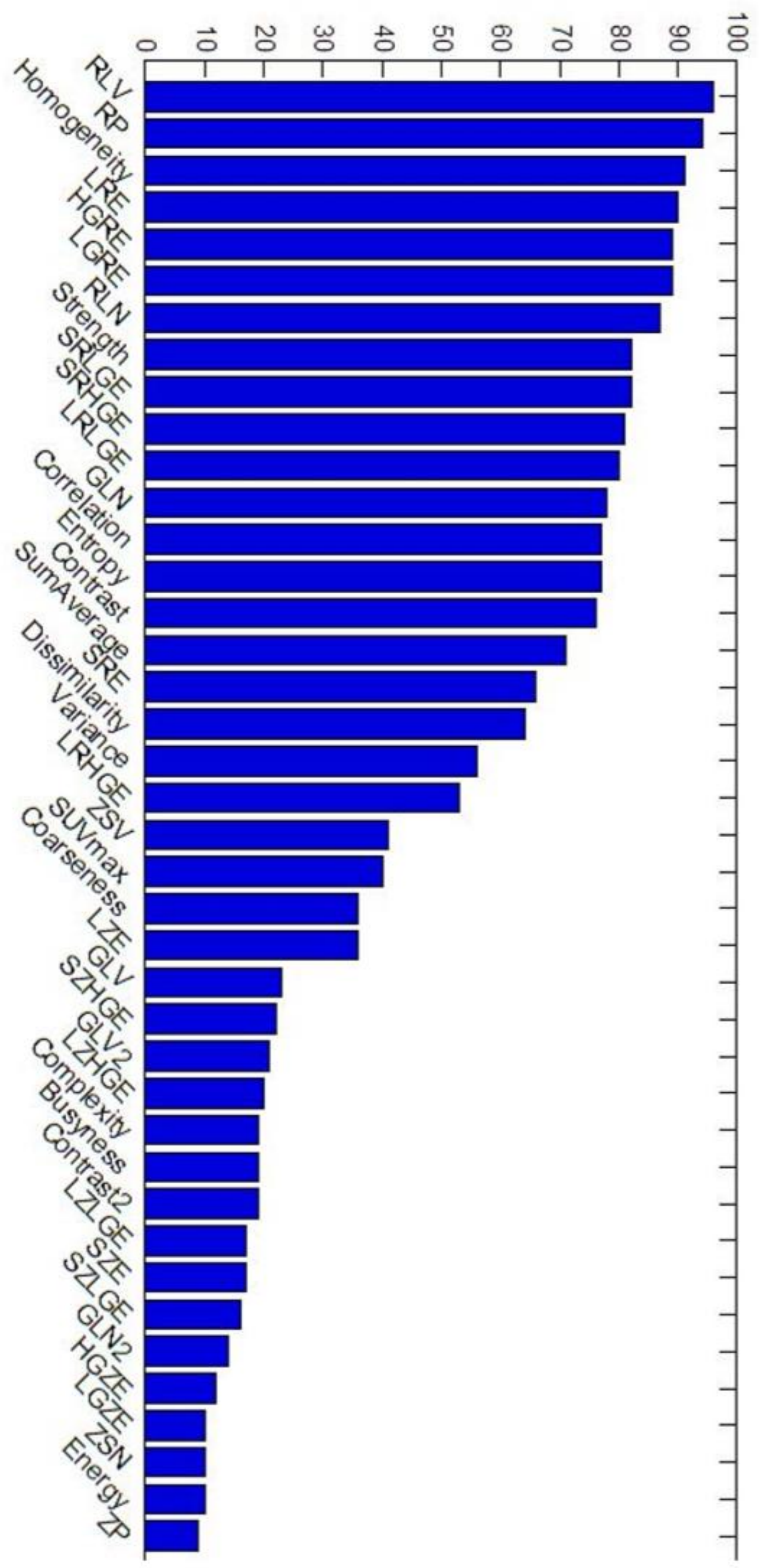


Figure 2.4 Selection frequencies of the features on training sets of the leave-one-out cross-validation when hybrid feature selection is employed, and data are normalized.

Chapter 3

Semantic Segmentation of PET images

3.1 Background Information

According to the Global Cancer Statistic report [73], lung cancer is the leading cause of death among the other cancers, with approximately 2 million deaths across 185 countries in 2020, and the second most diagnosed after breast cancer with 2 million cases. According to the American Cancer Society, Non-small cell lung cancer (NSCLC) is the most common subtype of lung cancer ranging from 80% to 85%. The biopsy is the golden standard for diagnosing NSCLC; however, imaging tests are frequently used with biopsy in detection, staging, assessment of therapy response, and prognostic evaluation.

Among the imaging tests, PET/CT is the most common modality compared to X-Ray and MRI due to the importance of radiation therapy planning and functional tumor volume assessment [74]. Positron emission tomography (PET) with 2-deoxy-2-[fluorine-18] fluoro-D-glucose (18F-FDG) provide functional information based on the radiolabeled glucose uptake in metabolically active tumors. FDG-PET is valuable for staging, restaging, radiotherapy planning, and biopsy guidance in oncology.

Manual segmentation is a time-consuming, tedious task in medical imaging. Furthermore, manual segmentation reproducibility is poor. Also, PET image resolution and SNR value is low compared to the CT and MRI modalities. The first automatization approach in segmentation is threshold-based. Since the quantification of PET image is SUV value, they took some percentage of SUV to decide tumor and background regions. However, the binary threshold is not a solution since there is no convention to which value should be used. Secondly, SUV values are affected by many biological and physical factors and are hard to standardize.

Over the years, many different techniques have been suggested based on optimization, statistic, etc. [75]. Hatt et al. [76] defined one of the important problems as no benchmark dataset to decide which approach provides better and fair comparison. First, most of the datasets used in publications are not open to public. Another problem is that only one expert draws the boundaries of the tumors. Furthermore, most publications ignore the repeatability, reproducibility, and robustness of segmentation based on scanner type and reconstruction parameters. All the arguments explain why tumor segmentation is still an active area of research.

Many institutions attacked the problem of creating a benchmark dataset for evaluating PET automatic segmentation algorithms for tumor delineation, such as the American Association of Physicists in Medicine (AAPM). Besides, creating dataset evaluation criteria is critical. AAPM created the first PET tumor segmentation challenge [77] to evaluate the online platform's state-of-the-art delineation algorithm. The Challenge dataset includes solid tumor PET images containing simulated, phantom, and clinical images. At the end of the challenge a CNN based model won the competition. This shows that higher-level feature extraction through the layers has a superior performance even in the small dataset.

Medical image segmentation consists of classifying each pixel or region that belongs to the organ, tumor, or structure. Radiologists visually inspect the images and define the boundaries manually or using an available software. Segmentation is a time-consuming and tedious task, and manual segmentation suffers from intra- and inter-variability.

Before deep learning gained attention, segmentation methods mainly were based on thresholding, region-based, boundary-based, stochastic, and learning-based [78]. Most of the deep learning-based medical segmentation methods in the literature are based on FCN [48], U-Net [49], 3D U-Net [79], V-Net [80] architectures. General segmentation strategy includes fine-tuning of the parameters specific to image dataset, modification of architecture, or changing cost function. The most applied problem is organ segmentation, and MRI is the leading modality due to its high image resolution [81].

This chapter introduces adaptation of DeepLab [82] architecture for lung tumor segmentation in FDG-PET images. We use Tversky loss for the class imbalance between large and small tumors [83]. We introduce a new PET based lung tumor dataset with the masks and conduct an experiment performance comparison of DeepLab with UNet.

3.2 Materials and Methods

3.2.1 Dataset

In this retrospective study, 141 patients with non-small cell lung cancer (NSCLC) that previously underwent 18F-FDG-PET/ CT imaging for cancer staging before surgery, chemotherapy, or radiotherapy treatment according to the stage of their disease from March 2010 to April 2014 were evaluated in the Acıbadem Kayseri Hospital. This study was approved by the research ethics committee of the Kayseri Research and Training Hospital. All procedures involving human participants' studies followed the institutional and/or national research committee's ethical standards and the 1964 Helsinki Declaration and its later amendments or comparable ethical standards. Informed consent was obtained from all individual participants included in the study.

3.2.2 Proposed Model

We adapted DeepLab version 3 (DeepLabV3) for semantic lung tumor segmentation. In [82], they use atrous (dilated) convolution instead of pooling and down sampling layers, which causes the loss of spatial information for a deeper network. Atrous convolution is a layer with a stride parameter that allows changing the filter’s field of view and carrying information to deeper blocks by inserting rate-1 zero to consecutive filter values. The main advantage is to extract denser features without extra parameters. Additionally, they introduce spatial pyramid pooling to capture multi-scale information. To capture different scale information, they use 4 different rated dilated convolutions in parallel and then concatenate the result. We used MobileNet-v2 [84] pretrained on ImageNet as a backbone to DeepLabV3. ASPP module was placed next to block 16 with rates (6,12,18) 3x3 convolutions as in [82]. We interchanged the classification layer with the Tversky loss layer with penalty terms $\alpha = 0.4$ and $\beta = 0.6$. The Tversky index helps improve the model's performance while the dataset is imbalanced. We used the Tversky layer for two reasons. The first reason was that the tumor size was small compared to the background. The second reason was the tumor size changed through slices (a need for regularization). It was also hard to convert 2D segmentation to 2.5D segmentation for further study.

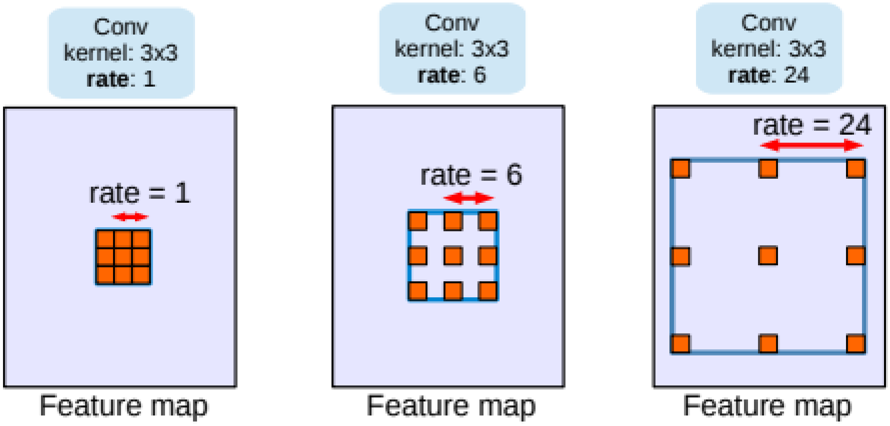


Figure 3. 1 Atrous convolution kernel [82]

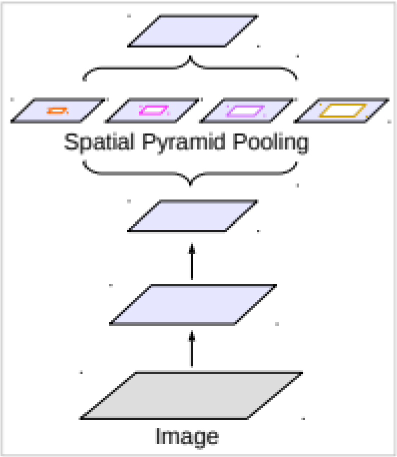


Figure 3. 2 Spatial pyramid pooling [82]

$$T(\alpha,\beta) = \frac{\sum_1^N p_{0_i} g_{0_i}}{\sum_1^N p_{0_i} g_{0_i} + \beta \sum_1^N p_{0_i} g_{1_i} + \alpha \sum_1^N p_{1_i} g_{0_i}} \qquad (3.1)$$

where α, β control the recall and precision tradeoff. P and G are predicted and ground truth images where p_{0_i} is the probability of pixel i being a tumor and p_{1_i} is the probability of pixel i being a background.

3.2.3 Experiment Design

We applied ten-fold cross-validation on our dataset; we trained with 131 patient images and tested for 10 images. We had such 10 training and test sets. Results were based on the average of these test sets. For each dataset prepared, we compared DeepLabV3 and U-Net architectures using three evaluation criteria whose details are given below.

In addition, we applied data augmentation by random rotation (-10 to 10) and reflection during each training phase. We performed Adam optimizer with a decay rate of 0.99, the initial learning rate of 0.0003 multiplied by 0.9 every 126 steps during one epoch, and a mini-batch size of 1. Three different cropping sizes were evaluated such as 32x32, 64x64 and 256x256. The same frameworks were investigated on cropped images with single tumor. In this study, the effects of augmentation, cropping and Tversky loss on the segmentation performance of the proposed model based on DeepLabV3 architecture were investigated. We implemented our codes in MATLAB.

3.2.4 Evaluation Metrics

The definitions performance metrics for intersection over union (IoU), Dice similarity index (DSI) and F1 score are as follows:

$$\mathrm{IoU} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \tag{3.2}$$

$$\mathrm{DSI} = \frac{2 * \mathrm{TP}}{2 * \mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \tag{3.3}$$

$$\mathrm{F1\ score} = \frac{2 * \mathrm{precision} * \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}} \tag{3.4}$$

where TF, FP, and FN are the true positives, false positives, and false negative rates. F1 score can also be mentioned as boundary F score or BFScore. Here we will use BFScore.

3.3 Results and Discussion

In Figure 3.3, the segmentation results are shown on one sample image for different frameworks such as U-Net, DeepLabV3, DeepLabV3 and Augmentation used together, and DeepLabV3 and Augmentation and Tversky Loss used together. The last framework gave the best segmentation performance both visually and quantitatively. The resultant evaluation metrics for each framework are listed in Table 3.1. Here, to highlight the advantage of using augmentation and Tversky loss along with DeepLabV3 architecture the image/slice that contains two bright spots is chosen. The center bright spot occurred due to the heart tissue, i.e., it does not indicate a tumor. As depicted in this figure, DeeplabV3 architecture is able to overcome multiple bright spots as opposed to the U-

Net architecture. In this study, only solid tumors (one tumor in each slide) were included in the analysis.

Table 3.1 demonstrates the comparative results of segmentation performance for DeepLabV3 and U-Net frameworks and the effect of augmentation and Tversky loss on the DeepLabV3 using three metrics such as mean IoU, BFScore and DSI. The scores are the average of all 10 test sets. Using augmentation with Tversky loss improves segmentation performance as can be seen from Table 3.1. Mean IoU scores are similar because tumor size is significantly small compared to the whole image. However, improvements can be observed even better using DSI and BFscore. Even though DSI values are similar in DeepLabV3 and U-Net architectures, BFscore values that take precision and recall into account are considerably different. It is evident that DeepLabV3 approach exhibits superior performance compared to the U-Net. It is worth noting that augmentation do not have considerably effect on the segmentation performance, however, Tversky loss layer increased DSI scores remarkably.

Figure 3.4 illustrates successful segmentation outcomes from six patients whose tumors were at different anatomical locations with different sizes using the DeepLabV3 and Tversky loss model with augmentation. As mentioned earlier, we worked on primer tumors, which assume each slice has one solid tumor. Even if there is another hot spot (bright mass) in the slice, the model can distinguish the difference and classify the second bright mass as a background. For instance, in the third row of Figure 3.4, there is a second mass on the image on the left, an artifact and a heart. In Figure 3.3, one can observe that a similar artifact could significantly affect the segmentation result. The Tversky loss layer helped to regularize artifacts in the image. Dilated convolution approach showed resistance to such artifacts compared to the U-Net. In the dataset, there are few examples of such artifacts. We expected a worse DSI score since a few examples are insufficient to learn.

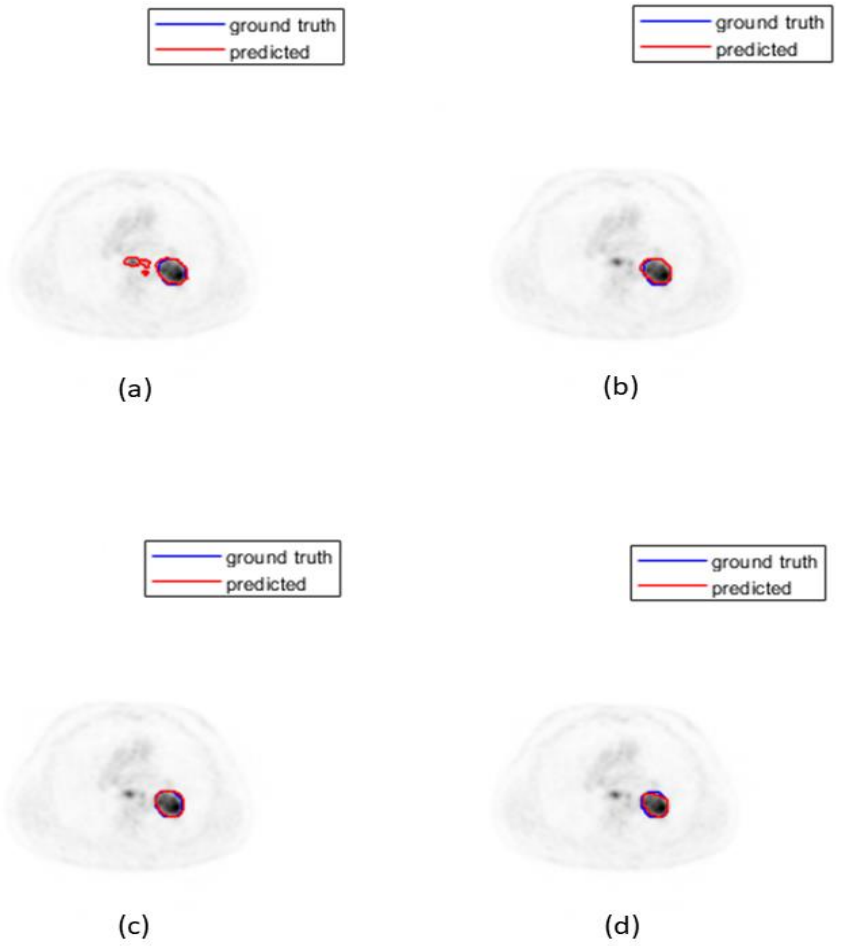


Figure 3. 3 Segmentation results for (a) U-Net, (b) DeepLabV3, (c) DeepLabV3 and Augmentation, and (d) DeepLabV3 and Augmentation and Tversky Loss. Blue and red lines indicate the ground truth and predicted segmentation outcomes respectively.

Table 3. 1 Segmentation performance comparison for different frameworks and augmentation and Tversky loss on the DeepLabV3 architecture.

	Mean IoU	**BFScore**	**DSI**
DeepLabV3	0.783504246	0.767433378	0.674516318
DeepLabV3 + Augmentation	0.784975141	0.78843336	0.68368549
DeepLabV3 + Augmentation + Tversky Loss	**0.809876166**	**0.79074866**	**0.735556089**
U-Net	0.757720767	0.62869407	0.645562299

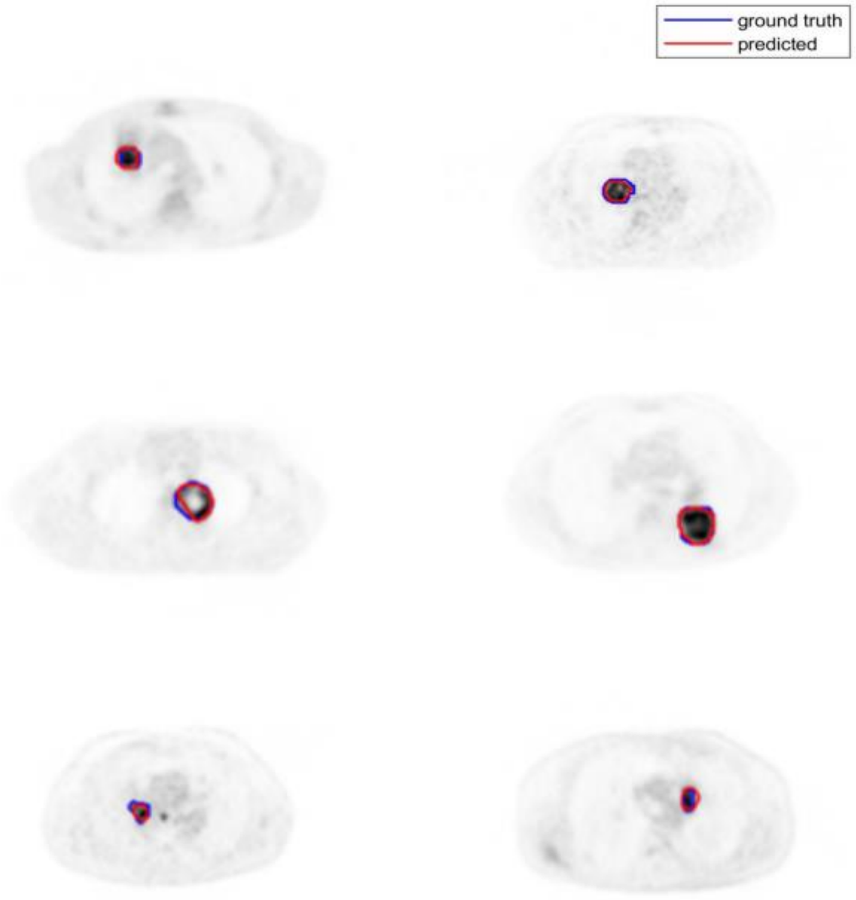
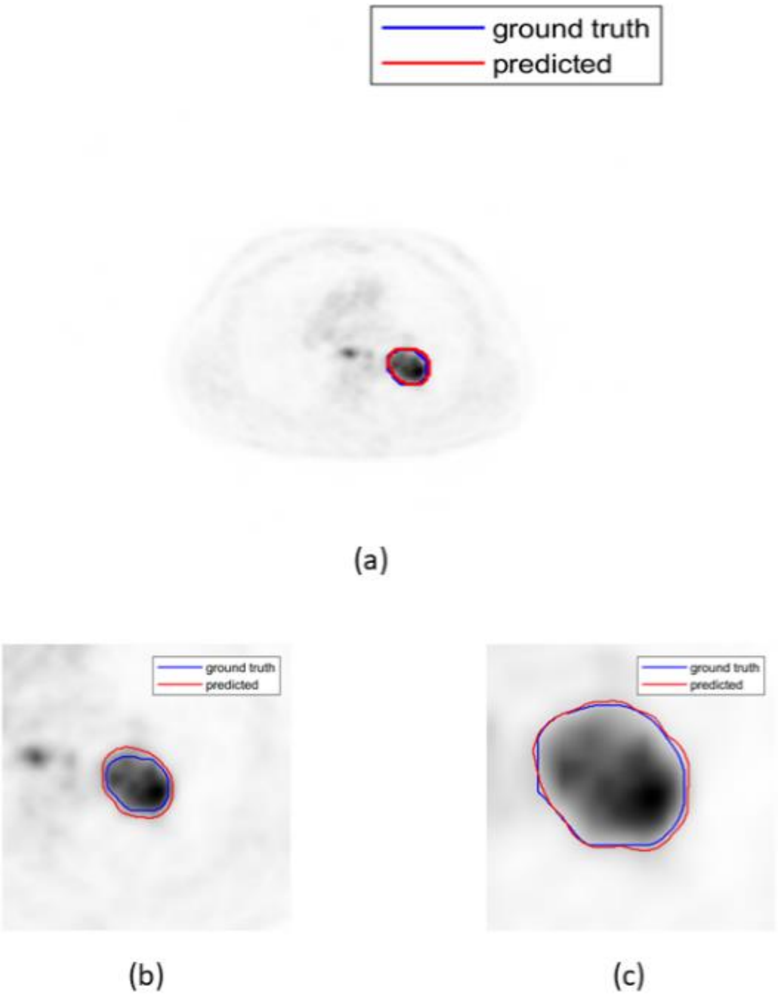


Figure 3. 4 Segmentation results for (a) U-Net, (b) DeepLabV3, (c) DeepLabV3 and Augmentation, and (d) DeepLabV3 and Augmentation and Tversky Loss. Blue and red lines indicate the ground truth and predicted segmentation outcomes respectively.



(a)

(b)

(c)

Figure 3. 5 Segmentation outcomes for cropped images with different box size such as (a) 256x256 (b) 64x64 (c) 32x32. Blue and red lines indicate the ground truth and predicted segmentation outcomes respectively.

Figure 3.5 demonstrates the segmentation performance of DeepLabV3 with Tversky loss and augmentation on one tumor with different box size, not the whole slice visually. Briefly, the segmentation performance was superior on 64x64 box size when compared to 32x32 in terms of BFScore (as shown in c). Both experiments showed that DeepLabV3 converges better compared to U-Net. Table 3.2 shows the cropped tumor segmentation results. As expected, cropping the tumor site improves segmentation results. However, 64x64 is better than 32x32 segmentation results. The main reason is that Tversky loss adds constraint and negatively affects the segmentation performance. Using DeepLabV3 without Tversky loss gives better results. The main reason for using DeepLabV3 with Tversky loss is to work on the whole image rather than the cropped versions since cropping the image will require adding an extra step to the procedure. Without cropping the tumor environment, satisfactory results can still be accomplished. One point to mention is that we created three different training sets for each fold. Each fold had the same patient ID and had different cropping sizes. For each cropping scheme, we created a new segmentation mask. This was because the interpolation step distorted the segmentation mask. We observed that if we used the ground truth mask in 256x256 size, cropped the tumor region (64x64, 32x32), and resized the input and masks to match the model's size, the mask boundaries were distorted. To prevent this, we created each size segmentation again with the help of the nuclear medicine doctor in our research team.

Table 3. 2 Segmentation results on cropped images using DeepLabV3 with Tversky loss and augmentation.

Cropping Box Size	**Mean IoU**	**BFScore**	**DSI**
256x256	0.809876166	**0.79074866**	0.735556089
64x64	**0.868627545**	0.588774377	0.832548396
32x32	0.848804223	0.271532421	**0.853158922**

Tables 3.3 to 3.5 show the performance of the proposed model with cropping of the tumor site. Cropping the image does not always improve the result. The tumor environment is also important. Our general observation is that the extra workload of cropping is removed in the proposed model.

If we look closer to the data, cropping tumor site significantly affects Patient 102. For several patients, cropping is necessary for better results. Cropping size being 64x64 or 32x32 does not substantially affect the Dice similarity index for this patient. For our dataset, 64x64 cropping is a better option since the tumor environment is clean and performance does not change significantly.

For each fold, we examined patient slices which yielded poor performance. In Figure 3.6, failures of segmentation are shown. This is the study's limitation since we have a small dataset compared to the patient's variability. Increasing the number of data would help to overcome these failures. Even for the same patient, slice-to-slice environment, pixel variation, and image quality were different.

Optimizing small datasets makes it harder to generalize the result; most of the time, it is impossible to produce one perfect solution. For this model and other deep learning research, the performance of the segmentation is limited to the dataset's quality.

That is why we opened the dataset for further research and contributed a multi-cancer research dataset.

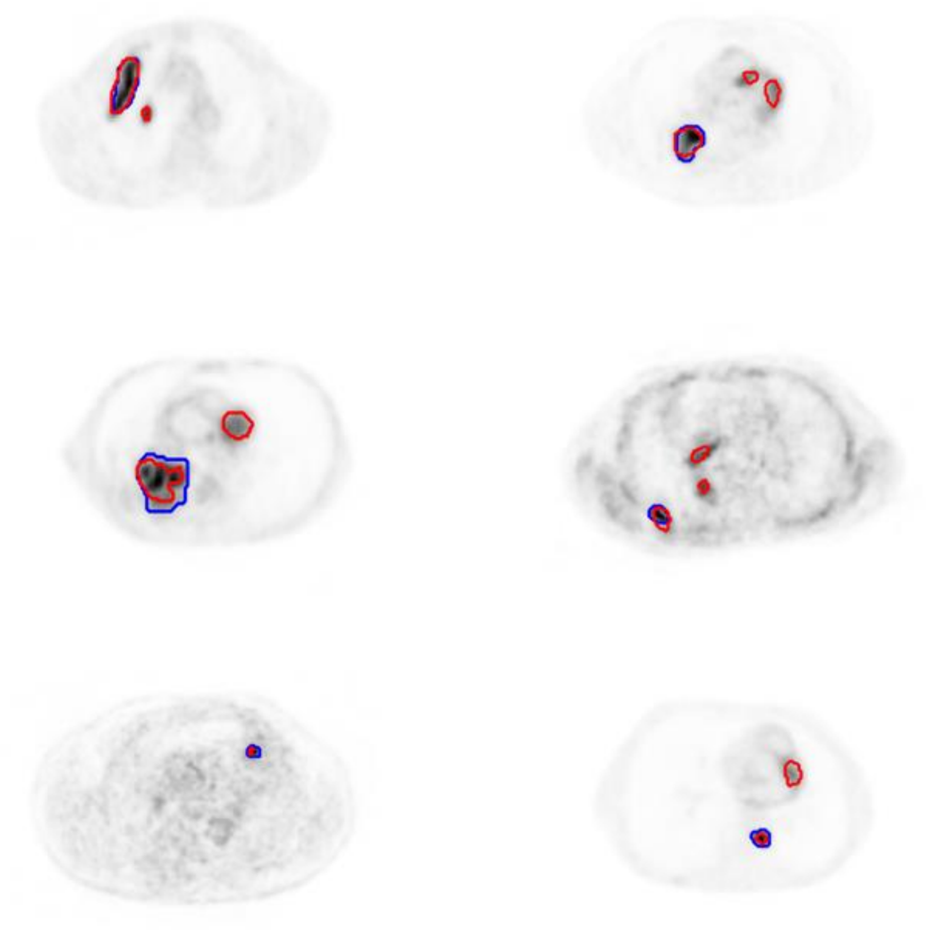


Figure 3. 6 Failure of proposed methods

In Figure 3.6, failures of the proposed segmentation results are shown. We have 131 patients with a variety of physical and metabolic conditions. Even though average scores are high, segmentation failures occurred on 6 patients. The reasons are that some of the tumors are too small, and these patients are outliers compared to the other patients in intensity variations. To overcome these failures, the proposed methods can be used as a preprocessor, and experts can choose the tumor or roughly crop the window to help segment the tumor.

The purpose was to automate the segmentation process. Doctors should always be in the loop during clinical applications. The model's overall performance is not enough to translate models to the clinic. For instance, one patient segmentation result did not improve overall performance; however, the result is significant for this patient as a human being. When we designed the model, we assumed final approval or modification applied by an expert. The proposed model accelerates the process of hand-crafted manual segmentation.

Table 3. 3 Using DeepLabV3 with Tversky loss and augmentation, Fold 6 of test images result (256x256)

Patient ID	**Mean IoU**	**BFScore**	**DSI**
102	0.621715	0.8125	0.39215
105	0.960477	1	0.95901
110	0.903009	0.91106	0.89302
115	0.661399	0.54321	0.48979
120	0.893036	1	0.88043
127	0.633745	0.53731	0.42424
46	0.859733	1	0.83720
64	0.912799	1	0.90476
67	0.710078	0.31772	0.59615
83	0.812202	0.95082	0.76923
92	0.89249	0.88495	0.88
94	0.859228	0.97029	0.83687
95	0.730436	0.61748	0.63325
96	0.824733	0.93333	0.78787
Avg	0.805363	0.81990	0.73457

Table 3. 4 Using DeepLabV3 with Tversky loss and augmentation, Fold 6 of test images result (64x64)

Patient ID	**Mean IoU**	**BFScore**	**DSI**
102	0.816499	0.44221	0.779046
105	0.952405	0.829703	0.953309
110	0.851851	0.540785	0.838471
115	0.709104	0.220779	0.604044
120	0.916518	0.760183	0.911594
127	0.952126	0.894231	0.950555

46	0.919871	0.689893	0.915497
64	0.921526	0.748655	0.919073
67	0.821339	0.252473	0.806873
83	0.910726	0.698124	0.904836
92	0.843744	0.536137	0.828524
94	0.933909	0.641462	0.933747
95	0.934965	0.741845	0.93387
96	0.937711	0.909382	0.935227
Average	0.887307	0.636133	0.872476

Table 3. 5 Using DeepLabV3 with Tversky loss and augmentation, Fold 6 of test images result (32x32)

Patient ID	**MeanIoU**	**BFScore**	**DSI**
102	0.8097	0.1206	0.7766
105	0.9423	0.5593	0.9540
110	0.8608	0.5447	0.8796
115	0.8186	0.0253	0.7960
120	0.9649	0.7773	0.9675
127	0.8450	0.2271	0.8316
46	0.9158	0.4249	0.9193
64	0.9190	0.4716	0.9290
67	0.6527	0.1116	0.7020
83	0.8059	0.1056	0.7918
92	0.7782	0.1514	0.7978
94	0.9183	0.3657	0.9332
95	0.7907	0.3347	0.7944
96	0.9032	0.4823	0.9026
Average	0.8518	0.3359	0.8554

Table 3. 6 Model without augmentation overall 10-fold result

	Acc	**mAcc**	**mIoU**	**wIoU**	**BFScore**	**Jaccard**	**DSI**
Fold1	0.9972	0.9854	0.7800	0.9963	0.8690	0.5628	0.6908
Fold2	0.9982	0.7015	0.6921	0.9966	0.8095	0.3860	0.5294
Fold3	0.9975	0.9203	0.7986	0.9963	0.8543	0.5996	0.7171

Fold4	0.9986	0.9266	0.8094	0.9980	0.9163	0.6201	0.7376
Fold5	0.9984	0.9425	0.8091	0.9975	0.9097	0.5012	0.6065
Fold6	0.9987	0.8674	0.7836	0.9980	0.9076	0.5684	0.6570
Fold7	0.9985	0.7791	0.7666	0.9970	0.9081	0.5348	0.6487
Fold8	0.9988	0.8386	0.8112	0.9977	0.9170	0.6236	0.7610
Fold9	0.9981	0.8994	0.7715	0.9974	0.8385	0.5448	0.6478
Fold10	0.9986	0.9052	0.8126	0.9977	0.9003	0.6266	0.7487
Average	**0.99831**	**0.8766**	**0.7835**	**0.9973**	**0.8830**	**0.5568**	**0.6745**

Table 3. 7 Model with Tversky loss 10-fold result

	Acc	**mAcc**	**mIoU**	**wIoU**	**BFScore**	**Jaccard**	**DSI**
Fold1	0.9982	0.95736	0.83332	0.9974	0.9035	0.6684	0.7677
Fold2	0.9985	0.93106	0.81529	0.9977	0.8985	0.6320	0.7335
Fold3	0.9966	0.97684	0.79629	0.9954	0.8407	0.5959	0.7093
Fold4	0.9973	0.94955	0.74497	0.9965	0.8316	0.4925	0.6178
Fold5	0.9987	0.90251	0.81987	0.9977	0.9133	0.6410	0.7702
Fold6	0.9988	0.87184	0.80536	0.9980	0.9050	0.6118	0.7345
Fold7	0.9986	0.96562	0.85195	0.9977	0.9239	0.7052	0.8203
Fold8	0.9978	0.98038	0.82154	0.9969	0.8830	0.6452	0.7543
Fold9	0.9983	0.93177	0.79009	0.9976	0.8776	0.5818	0.6883
Fold10	0.9986	0.95098	0.82004	0.9978	0.8993	0.6414	0.7592
Average	**0.9981**	**0.94179**	**0.8098**	**0.9973**	**0.8876**	**0.6215**	**0.7355**

Table 3. 8 Model with augmentation 10-fold result

	Acc	**mAcc**	**mIoU**	**wIoU**	**BFScore**	**Jaccard**	**DSI**
Fold1	0.9980	0.9114	0.7916	0.9971	0.8957	0.5852	0.7067
Fold2	0.9983	0.9430	0.8117	0.9976	0.8962	0.6251	0.7277
Fold3	0.9977	0.7634	0.7291	0.9957	0.8026	0.4605	0.6097
Fold4	0.9989	0.7728	0.7209	0.9981	0.9143	0.4429	0.5378
Fold5	0.9990	0.8231	0.7867	0.9981	0.9322	0.5745	0.6870
Fold6	0.9987	0.7650	0.7322	0.9975	0.8898	0.4658	0.5903
Fold7	0.9988	0.9519	0.8766	0.9980	0.9626	0.7544	0.8541
Fold8	0.9990	0.9039	0.8473	0.9982	0.9341	0.6956	0.8053
Fold9	0.9971	0.8941	0.7338	0.9962	0.8027	0.4706	0.5844
Fold10	0.9990	0.8308	0.8192	0.9982	0.8923	0.6394	0.7334
Average	**0.9984**	**0.8559**	**0.7849**	**0.9975**	**0.8922**	**0.5714**	**0.6836**

Table 3. 9 U-Net 10-fold result

	Acc	**mAcc**	**mIoU**	**wIoU**	**BFScore**	**Jaccard**	**DSI**
Fold1	0.9983	0.9764	0.7950	0.9974	0.8443	0.5917	0.7236
Fold2	0.9931	0.9887	0.6919	0.9917	0.6331	0.3908	0.5119
Fold3	0.9984	0.9183	0.8264	0.9972	0.8722	0.6544	0.7790

Fold4	0.9992	0.8905	0.8330	0.9985	0.9547	0.6667	0.7919
Fold5	0.9978	0.9051	0.7674	0.9967	0.8064	0.5370	0.6669
Fold6	0.9918	0.9704	0.6478	0.9903	0.6010	0.3038	0.4288
Fold7	0.9989	0.9780	0.8711	0.9982	0.9218	0.7433	0.8475
Fold8	0.9976	0.8170	0.7237	0.9960	0.7734	0.4497	0.6103
Fold9	0.9958	0.9575	0.7021	0.9948	0.6943	0.4085	0.5332
Fold10	0.9936	0.9898	0.7184	0.9923	0.7058	0.4432	0.5618
Average	**0.9964**	**0.9392**	**0.7577**	**0.9953**	**0.7807**	**0.5189**	**0.6455**

Table 3. 10 Comparison result for 256x256

256x256	**Acc**	**mAcc**	**mIoU**	**wIoU**	**BFScore**	**Jaccard**	**DSI**
DeepLabV3	0.9983	0.8766	0.7835	0.9973	0.8830	0.5568	0.6745
DeepLabV3 + Aug	**0.9984**	0.8559	0.7849	**0.9975**	**0.8922**	0.5714	0.6836
DeepLabV3 +Aug +Tver	0.9981	**0.9417**	**0.8098**	0.9973	0.8876	**0.6215**	**0.7355**
U-Net	0.9964	0.9392	0.7577	0.9953	0.7807	0.5189	0.6455

Table 3. 11 Comparison result based on image sized proceed

DeepLab+Aug+Tver	**Acc**	**mAcc**	**mIoU**	**wIoU**	**BFScore**	**Jaccard**	**DSI**
256x256	**0.998173**	**0.941798**	0.80987	**0.9973**	**0.8876**	0.6215	0.7355
64x64	0.989377	0.929216	**0.86862**	0.9808	0.7686	0.7483	0.8325
32x32	0.953838	0.938877	0.848804	0.9178	0.5420	**0.7541**	**0.8531**

In this study we proposed the use of DeepLabV3 architecture to segment non-small-cell lung tumors on FDG-PET images, especially on the slice with the SUVmax. In our previous study [85], we used random walk segmentation approach with a similar goal in which the user/physician defined two seed points one inside and one outside of the tumor region. However, this kind of an interactive segmentation is time consuming and requires user interaction and post-processing. However, the proposed model does not need any supervision and can successfully segment the tumor from the whole slice or the bounding box.

We compared the proposed model with the U-Net architecture which is highly popular in medical image segmentation. U-Net performed relatively well depending on the tumor size and location. However, for small tumors it either overestimated or underestimated the tumor boundary. DeepLabV3 solved this problem with atrous convolution layer and spatial pyramid approach. Further improvement came with the Tversky loss layer added to the MobileNet architecture. In most datasets including ours, small-size tumors are fewer than large ones which creates a significant imbalance in the training phase of the segmentation. We selected Tversky loss layer to refine tumor boundaries by penalizing incorrect segmentation. Moreover, augmentation during the training phase improved the segmentation performance which can especially be seen in dice similarity index.

As suggested in [86], for reproducibility we share our dataset to address limited annotated data problem in medical images. Cross-validation was applied to prevent overfitting problem and we used evaluation metrics of segmentation. The generalizability of the proposed solution is a challenging issue due to variability of patient and the limited number of data. However, contributing a new PET image dataset coming from lung cancer patients will add value to comparing workable solutions with different scanners and cancer types.

Further improvement can be accomplished with interactive and image-specific fine-tuning after applying DL-based segmentation as proposed in [87]. In this study, our aim was to decrease the time that the clinical expert spends time on manual drawing. The segmentation accuracy that is obtained in this study is acceptable by clinical experts. However, we are aware of the fact that there is inevitable variability among their decisions.

The proposed solution's purpose is to improve the accuracy of segmentation; in another work [89], we observed classification problem variation comes mainly from the classification method rather than the segmentation mask. Choosing 2D segmentation is not mandatory; the proposed solution can be converted to 2.5D segmentation, which applies to each patient slide and then converted to isotropic volume data using scanner parameters. Compared to the 3D methods 2D and 2.5D methods are cost-efficient [49].

In this paper, we proposed a semantic segmentation model for PET images of NSCL cancer. We compared the proposed method with U-Net and showed that Tversky loss and the proposed network significantly improved segmentation results. For reproducibility and generalizability we shared images with masks.

In recent years PET/CT multi-modality segmentation approach has been widespread. The idea behind the multi-modality segmentation is to use each modality’s advantage to improve segmentation accuracy and visual fusion map. However, we think that this is highly challenging for two reasons: First, the image resolutions and size of image obtained using each modality are different. In our case, the image size of PET is 168x168. On the other hand, CT image size is 256x256. If you want to use two modalities for the same region of interest, you need to interpolate or down-sample one of the modalities. For both cases, still, anatomical matching is required; registration of PET to CT is not a trivial task. Since the resolution of PET is low, it is hard to verify anatomical matching. In studies, the registration step was generally missing; they interpolated PET image size to CT. This process gave saliency maps. The second problem with multi-modality segmentation is that each modality generates a different tumor volume. CT based tumor volume is more problematic because functional features are not represented in CT. That is why the clinic uses the PET/CT modality, combining anatomical and functional imaging.

Commercial PET/CT devices register and show fused images to clinicians to segment tumors. Vendors generate the volume of scan CT and PET in DICOM images separately; they do not provide fused volume. This causes two sets of images with distinctive features and ground truth masks. Using two different modalities with different masks for one ground truth segmentation makes the assessment unclear. Another point is the problem definition. If we apply semantic segmentation to a classification problem, the segmentation solution is vital but decisive performance. The performance limitation occurs due to the classification problem (TNM staging, subtype classification, therapy response), classifier, and feature selection. However, when the problem is to decide the radiation dose for the therapy, segmentation performance may affect the final outcome. Even though there are such difficulties, multi-modality segmentation studies provided valuable information for semantic segmentation of lung tumors in PET images.

In [88], 3D-UNet was applied for PET segmentation for extracted patches. The Dice score is 0.85 for PET-only segmentation without cross-validation for 60 NSCLC cases. They compared their method with a graph-based segmentation algorithm and found deep learning approach outperforms to segmentation bot PET and CT images. They found multimodality feature fusion had limited improvement on PET-only segmentation. 3D U-Net performance for PET images was under 70%, similar to our dataset. In [89], DSI is 0.86 for segmentation using patches/boxes without cross-validation. In [90], DSI is 0.85 for multi-modality segmentation for patches without cross-validation for 84 NSCLC patients.

Using patches/boxes increases the DSI significantly; however, it is unrealistic to ask doctors to create a bounding box for each slice. Because a data-driven approach is a power supported by data, models should be built upon more available data. Graph-based and parametric segmentation methods perform well with preprocessing and post-processing approaches. Dilated convolution approach with pyramid pooling performs well without a bounding box. The article above also suggested that a well-structured natural image segmentation approach will improve the performance. The main limitation in medical semantic segmentation is annotated data. It is impossible to compare each deep learning state-of-the-art segmentation method. While designing an experiment, we expected that DeepLabV3 will outperform U-Net because of the technical contribution of dilated convolution, pyramid pooling, and MobileNet backbone. We compared U-Net since it is very popular and a general improvement modifying it according to a specific problem.

As mentioned in [51], many possible improvements can be made to improve the segmentation performance. One is architecture level; we borrowed DeepLabV3 from natural image processing. Another is the loss function; we used Dice loss which is more suitable compared to the cross-entropy loss function, and finally, we used the Tversky loss layer to solve the imbalanced tumor size problem.

Chapter 4

Prognostic Value of Radiomics in Pancreatic Cancer

4.1 Introduction

Pancreatic adenocarcinoma (PA) is one of the mortal cancers with, a five-year survival rate is 3% for distant tumors based on SEER staging [91]. The common treatment of PA is neoadjuvant chemotherapy (NC) which chemotherapy drugs are administered before undergo resection of the tumor. If resection is not possible although the NC, radiation therapy can be used for possibility of surgery. However, a major problem with this treatment is only 20% of tumors are able to the resection. [92].

FDG-PET clinical application was explained in chapter one. 18F-FDG PET/CT is an accurate and useful modality in PA diagnosing, staging and treatment response [93]. According to meta-analysis study FDG PET/CT diagnosing PA sensitivity and specificity can be reached 90% and 80% respectively [94].

One of the most common clinical features to assess treatment response is SUV value. Recently researchers have shown higher SUV value is associated with prognosis of PA [95]. Alternative studies MTV and TLG [96] can be used as a prognostic factor beside SUV.

However, these clinical features are having serious problem with standardization issue because modality parameters affect the values of SUV, MTV and TLG. Addition to variation in imaging parameters, inflammatory lesion and small volume of the tumor affect these clinical feature values.The past decade has seen the rapid development of radiomics approach, image-derived features, in quantification of tumor heterogeneity. Radiomics showed association with survival in pancreas patient [97]. In the present study, we evaluated radiomics features extracted from 18F-FDG PET images for predicting OS in patients with PA.

4.2 Materials and methods

We retrospectively reviewed the electronic medical records of 72 patients who were histopathological diagnosed with pancreatic adenocarcinoma in Başkent University, Adana, Dr. Turgut Noyan Application and Research Center Hospital between March 2006 and January 2017. Patients who received neoadjuvant chemotherapy, radiotherapy, surgical tumor resection, stent, and drainage catheter were excluded. A total of 72 patients (37 men and 35 women) were eligible and included in the study. A complete demographic description of the patient population is shown in Table 4.1.

Table 4. 1 Demographic and clinical characteristics of the 72 patients

Characteristic	Value
Age	64.9±10.3
Gender, n (%)	
Female	35 (48.6)
Male	37 (51.4)
Tumor size (cm),	4.6 ± 1.4 (1.9 – 8.6)*
Tumor maximum SUV	9.5 ± 4.8 (3.8 – 29.1)*
Metabolic tumor volume (cm3)	44.7 ± 41.3 (2.54-176.5)*
Clinical stage, n (%)	
I	10 (13.9)
II	13 (18.1)
III	11 (15.3)
IV	38 (52.8)

*mean ± SD (range)

18F-FDG PET/CT images were acquired an integrated scanner (Discovery-STE 8; General Electric Medical System, Milwaukee, WI, USA). All patients were instructed to fast for at least 6 hours before the intravenous administration of 5 MBq/Kg ^{18}FDG. We measured pre-injection blood glucose levels to be sure that it is below than 200 mg/dL. Approximately 60 min after the intravenous administration of FDG, an unenhanced CT scan with a slice thickness of 3.3 mm from the skull's vertex or base to the inferior pelvis's inferior border was performed at 80mA and 140kV. The subsequent PET scan was performed for 3 min in each seven-bed position under the three-dimensional mode from the vertex or base of the skull to the inferior border of the pelvis. FDG PET images were reconstructed using CT image for attenuation correction.

An experienced nuclear medicine specialist used PT/CT modality software to delineate tumor region. SUV, MTV and TLG was calculated from VOI. Regional lymph nodes were excluded in the VOI. For the radiomics features experts decided the range of slices for the tumor of each patient. Images were transferred from DICOM to PC using MATLAB. A tumor mask was extracted for each slice based on expert drawings. We used a random walk algorithm to extract tumor masks in MATLAB, manual adjustment was done based on expert visual inspection. All images were normalized to 0 to 255. Image plane resolution was 5.46x3.47 mm^2, and slice resolution was 3.27 mm. We isotopically resampled the volume to 3.27x3.27x3.27 mm^3. Uniform quantization was applied to each patient volume, and level-64 was chosen to compute image features. We extracted GLCM (8 features), GLRLM (13 features), GSZM (13 features), and NGTDM (5 features) from tumor region.

All statistical analyses were performed using the Statistical Package for the Social Sciences (SPSS for Windows, version 22) software program (IBM, Armonk, New York, USA). Continuous variables were expressed as mean ± standard deviation, and categorical variables were expressed as frequency (percentage). The statistical significance level was selected as $p < 0.05$, and all tests were two-sided. Confidence intervals (CIs) are reported at the 95% level. The primary endpoint was overall survival (OS) measured from the date of the PET/CT scan to the date of death from any cause or the date of the last clinical follow-up. To assess and compare the predictive performance of PET imaging parameters, we used time-dependent receiver operating characteristic (ROC) curves for censored survival data and areas under the ROC curve (AUC) two years after diagnosis. Univariate and multivariate analyses using Cox proportional hazards regression were performed to assess the relationship between PET imaging parameters and OS. Multivariable analysis adjusted for age, sex, clinical stage, and tumor size was performed. Kaplan-Meier curves were generated with an optimal cut-off value derived from maximally selected rank statistics for imaging parameters.

4.3 Results and Discussion

Time-dependent ROC curve analysis gave AUCs for 2-year survival prediction showed in Table 4.2. PET textural features and conventional PET indices were ranked based on predictive performance. First-order energy (AUC = 78.36) as a textural feature had the highest performance. The textural features with the highest performance after energy are strength (AUC = 77.64) and entropy (AUC = 77.25). Among the conventional PET parameters, maximum SUV (AUC = 68.33) and mean SUV (AUC = 68.33) gave the best index followed by mean SUV (AUC = 0.628), followed by MTV (AUC = 59.36) and

TLG (AUC = 51.49). The highest-ranking features in the time-dependent ROC curve analysis were selected for further analysis. Energy, strength, and entropy as textural features and clinical variables of age, sex, clinical stage, and tumor size were used as variables in Cox regression models.

Table 4. 2 Time-dependent ROC curve analysis for 2-year overall survival prediction

Textural features	AUC	95% CI
Energy	78.36	60.97-95.74
Strength	77.64	57.52-97.76
Entropy	77.25	58.13-96.36
Complexity	76.77	56.71-96.83
SZLGE	75.46	56.96-93.96
Coarseness	74.76	54.04-95.49
LGZE	74.07	54.39-93.75
GLN	73.76	50.42-97.11
LZLGE	73.76	55.32-92.20
Contrast.1	73.20	51.55-94.85
GLV	73.19	55.58-90.81
LRLGE	73.11	52.47-93.74
LGRE	73.03	52.80-93.25
SRLGE	73.03	52.80-93.25
ZSV	70.18	43.73-96.64
GLN.1	70.01	45.05-94.97
RLV	69.23	46.45-92.01
GLV.1	68.45	43.27-93.64
Maximum SUV	68.33	46.52-90.14
Mean SUV	68.33	46.52-90.14
Variance	67.44	48.69-86.20
MTV	59.36	33.14-85.59
SRHGE	57.83	31.08-84.59
HGZE	57.32	31.16-83.48
Busyness	57.27	35.06-79.48
HGRE	57.06	29.73-84.39
LRHGE	56.48	28.61-84.36
SZHGE	56.31	33.20-79.43
LZE	56.15	28.52-83.78
ZP	55.99	27.28-84.70
Sum Average	55.10	28.58-81.61
ZSN	54.91	24.88-84.93
SZE	54.63	24.37-84.90
LRE	54.50	26.95-82.06
RP	54.50	26.95-82.06
SRE	54.37	26.72-82.01
RLN	54.37	26.72-82.01
Correlation	54.29	23.77-84.81
LZHGE	53.26	23.27-83.25
Contrast	52.23	22.01-82.44
TLG	51.49	23.36-79.62
Dissimilarity	51.48	21.02-81.94
Homogeneity	50.07	17.49-82.65

In the univariable analyses, tumor size, energy, entropy, and strength were significant predictors of OS showed in Table 4.3. After adjusting for age and strength, the multivariable Cox analysis demonstrated that strength (hazard ratio, HR, 0.98 95 % CI 0.97 – 0.99 P = 0.005) was independently associated with better overall survival.

Kaplan-Meier analysis of the entire cohort demonstrated significantly improved survival in patients with higher strength tumors. Patients with lower strength tumors had a significantly shorter 2-year OS than those with higher strength tumors.

Table 4. 3 Univariate and Multiple Cox regression analysis results in identifying the risk factors of overall survival

Variable	Univariate		Multivariate	
	HR (95% CI)	*p*	HR (95% CI)	*p*
Age (years)	1.03(1.01-1.05)	0.043	1.02(1.01-1.05)	0.049
Gender (male/female)	1.09(0.64-1.85)	0.762	-	-
Clinical Stage				
I	1.00	-	-	-
II	1.88(0.71-4.97)	0.202	-	-
III	0.78(0.26-2.34)	0.655	-	-
IV	1.79(0.79-4.06)	0.164	-	-
Tumor size (cm)	1.17(1.01-1.36)	0.045	-	-
Energy (x1000)	0.64(0.47-0.90)	0.010	-	-
Strength	0.98(0.97-0.99)	0.003	0.98(0.97-0.99)	0.005
Entropy	1.66(1.15-2.40)	0.007	-	-

This study demonstrated that heterogeneity of 18F-FDG uptake measured by PET radiomics was an independent prognostic factor for survival in patients with PA. The strength of the primary tumor had a better prognostic value than metabolic parameters. Higher-strength tumors as a measure of heterogeneity were independently associated with more prolonged survival.

Chapter 5

Conclusions and Future Prospects

5.1 Conclusions

The primary aim of this study was to find the feature or features that can be obtained from the image to be used clinically. This is important because the feature used in classification has clinical problems and requires biopsy for definitive diagnosis. As a result, this thesis used machine learning techniques for lung cancer subtype classification.

Significant amount of time was spent to extract the features, collect patient data, and identify tumor sites of interest. An isotropic volume was created, and various features were extracted under the programming and expert physician consultation. Extracted features were classified using machine learning techniques. The fact that the features extracted from the image have an essential role in cancer studies. This is very important in terms of saving patients from biopsy and evaluating the course of their disease and their response to treatment.

The first part of this study is important in that it shows that similarly extracted features can be helpful in the diagnosis and that machine learning models will contribute to clinical applications. Some factors limit this study. The first is the limited number of patients. Although there is not a large enough data for a general solution, the number of patients is more than the number of articles in the literature. The complex metabolism of cancer, the differences in the scanning scanner, and the analysis methods are among the factors that make it difficult to evaluate the solution found. In addition, the unique nature of each cancer type has made it very difficult to find a feature or features that can be applied to the general clinic. For this reason, only lung cancer was focused on, and subtype discrimination was chosen.

While machine learning brings together statistical and mathematical models, since data production has increased exponentially, models to explain phenomena from data, deep learning, have become very popular. The success of models that feature the data as a whole rather than extracting specific features is increasing rapidly in image processing. It has also increased success using deep learning on the same dataset.

The second study's subject is determining the tumor region by the model using a deep learning approach. One of the problems we observed in the first study is that it takes the physician a long time to draw the area manually. Although there is a feature for this in the program of the PET/CT imaging device, it cannot be exported to be studied further on a PC for example. This makes the study difficult.

This problem is one of the most frequently studied problems in image processing. Finding a well-reconciled algorithm or model is difficult due to the organ studied, the extraction technique, and the disease. Our study has also several limitations in this context. The proposed model was developed to increase the success of our patient set. However, it is possible to train the model with more data and generalize its success in deep learning approaches. It is possible to see this in computer vision applications. It is possible to see the success of models based on data when there is enough data in smart homes, face and voice recognition programs, and autonomous vehicles. There are assorted reasons why applications in the health field do not develop at this speed. First, patient ethics is a relatively slow-developing field due to the difficulty of data availability and the economic value of the devices' software. In addition, each patient is far from being just data in health. The failure of a new model may result in the loss of human life, a responsibility that no one wants to take.

Especially in cancer research, doctors do not want to take responsibility for ethical reasons. In this case, studies accumulate in the literature but cannot turn into clinical applications. Although each study's limitations depend on its data, it is important to emphasize the probes and provide information.

Our second study developed a model by considering these issues and successfully determining the tumor region for this dataset. Sharing the dataset we used aims to benefit those working in this field and provide different patient data in the field of cancer-related imaging.

For this, a deep learning approach has been applied. As mentioned above, learning the system over data has helped solve many problems. This is true for semantic segmentation as well as other classification problems. Similarly, deep learning model pixels can be trained whether they are part of the object. We also benefited from these approaches in PET images of lung cancer patients. One of the problems with our dataset was the limited number of small tumors. This causes the system trained with large tumors to choose a larger area than it does for such problems. Thus, we wanted to reduce the selected areas that are larger than they are. For this, a new constraint equation is needed when the augmentation methods are insufficient. At this point, we have benefited from the Tversky equation.

This study aimed to save time for physicians on tumor segmentation and provide better results as data is added to the system. We achieved approximately 90% success in our data by adapting the definitive version of the model to the semantic segmentation problem with the DeepLab method of MobileNet architecture and using the Tversky equation. We compared this method with another popular method such as U-Net.

When evaluated together with our first study, we can segment the tumor autonomously and successfully classify it by extracting features from it. As the number of data increases, the models learn better, and it is expected that the data and studies in this field will increase. These approaches will take their place in the clinic as a complementary tool, not a substitute for doctors, in the processes from the first shooting to diagnosis, then treatment planning and follow-up of the disease. We can think of this thesis as an assistive system design for doctors. We hope that because of cancer studies, humanity will find the cure for cancer, and scientists will focus on studies on the benefits of artificial intelligence on different subjects.

5.2 Societal Impact and Contribution to Global

This thesis contributes to the effort to improve classification and prognosis performance. In 2018, international cancer research funding reached USD 5.5 billion [98]. While treatment research is the top category with 25%, the diagnosis and prognosis category are approximately 15%.

Global oncology spending was USD 126 billion in 2018. Besides the economic damage, more importantly, we lose people. The National Cancer Institute estimates that 130,180 people died due to lung cancer, 21% among the cancer types, in 2022, and newly diagnosed lung cancer is 236,740. We prepared an open database for lung cancer. It will serve as another benchmark for further investigation. Open datasets in the field are highly limited.

5.3 Prospects

5.3.1 Software as a Medical Device (SAMD)

The Food and Drug Administration (FDA) published a regulation and action plan for Artificial Intelligence & Machine Learning based software [99]. Due to an increased amount of software in medical application and their success, regulation and validation are required before clinical usage. The critical point is that the FDA accepts software as a medical device even though there is no hardware part. There are four hundred software packages approved and listed on the official website. Approximately 75% of SAMD is related to radiology, and several of them are developed for PET devices.

The main aims of the action plan are to encourage good machine learning practice through the software (bias elimination, robustness, etc.) and standardization. According to the Guideline for Clinical Evaluation [100], our work is in the early clinical association stage. In the future, our purpose is to add more data to generalize the output and develop more reliable and precise software to be listed in SAMD.

5.3.2 Interpretability and Explainable Machine Learning

Interpretability of the model can be defined as explaining how a particular model decides specific outputs. It is a crucial topic; especially ML/DL models are used in bank credit, law, and the healthcare system. Model interpretability is a new topic in the literature compared to ML and DL. It is because people prefer accuracy over understanding how models decide certain outputs.

Another challenge is to measure whether the model is explainable or interpretable. There is no consensus and standard in this field. While evaluating interpretability, we have global and local tools [98]. Global tools focus on understanding the reasoning leading to all possible outcomes. On the other hand, local tools focus on understanding specific decisions.

Evaluating explanations can be categorized as application-grounded, human grounded, or functionally grounded. In our case, our model should be application grounded so that we can explain our model to domain experts (doctors). In the future, we will develop an interpretable and explainable model reviewed in [101].

5.3.3 Cloud and Tensor Processing Unit

In this thesis, we used a workstation and personal computer since the patient population is limited to two hundred patients. However, a clinical application needs millions of patients for validation. Cloud platforms are essential for sharing multicancer data and computation performance. On the other hand, hardware technology is improving. GPU accelerated ML and DL research; now we have TPUs [102]. In the future, we will move to cloud service and write programs for TPUs.

BIBLIOGRAPHY

[1] G. Omami, D. Tamimi, and B. F. Branstetter, “Basic principles and applications of 18F-FDG-PET/CT in oral and maxillofacial imaging: A pictorial essay,” *Imaging Sci. Dent.*, vol. 44, no. 4, pp. 325–332, 2014.

[2] D. R. Schaart, “Physics and technology of time-of-flight PET detectors,” *Phys. Med. Biol.*, vol. 66, no. 9, 2021.

[3] A. Gallamini, C. Zwarthoed, and A. Borra, “Positron emission tomography (PET) in oncology,” *Cancers (Basel).*, vol. 6, no. 4, pp. 1821–1889, 2014.

[4] C. Gridelli *et al.*, “Non-small-cell lung cancer,” *Nat. Rev. Dis. Prim.*, vol. 1, pp. 1–16, 2015.

[5] F. C. Detterbeck, D. J. Boffa, and L. T. Tanoue, “The new lung cancer staging system,” *Chest*, vol. 136, no. 1, pp. 260–271, 2009.

[6] V. K. Anagnostou *et al.*, “Molecular classification of nonsmall cell lung cancer using a 4-protein quantitative assay,” *Cancer*, vol. 118, no. 6, pp. 1607–1618, 2012.

[7] I. Dagogo-Jack and A. T. Shaw, “Tumour heterogeneity and resistance to cancer therapies,” *Nat. Rev. Clin. Oncol.*, vol. 15, no. 2, pp. 81–94, 2018.

[8] Y. Balagurunathan *et al.*, “Reproducibility and Prognosis of Quantitative Features Extracted from CT Images.,” *Transl. Oncol.*, vol. 7, no. 1, pp. 72–87, 2014.

[9] L. Alic, W. J. Niessen, and J. F. Veenland, “Quantification of heterogeneity as a biomarker in tumor imaging: A systematic review,” *PLoS One*, vol. 9, no. 10, pp. 1–15, 2014.

[10] R. Samala, W. Moreno, Y. You, and W. Qian, “A Novel Approach to Nodule Feature Optimization on Thin Section Thoracic CT,” *Acad. Radiol.*, vol. 16, no. 4, pp. 418–427, 2009.

[11] F. Tixier, M. Hatt, C. Cheze-Le Rest, A. Le Pogam, L. Corcos, and D. Visvikis, “Reproducibility of Tumor Uptake Heterogeneity Characterization Through Textural Feature Analysis in 18F-FDG PET,” *J. Nucl. Med.*, vol. 53, no. 5, pp. 693–700, 2012.

[12] I. El Naqa, “The role of quantitative PET in predicting cancer treatment outcomes,” *Clin. Transl. Imaging*, vol. 2, no. 4, pp. 305–320, 2014.

[13] M. Sollini *et al.*, “PET Radiomics in NSCLC: state of the art and a proposal for harmonization of methodology,” *Sci. Rep.*, vol. 44, no. 1, p. 358, 2016.

[14] P. E. Galavis, C. Hollensen, N. Jallow, B. Paliwal, and R. Jeraj, “Variability of textural features in FDG PET images due to different acquisition modes and reconstruction parameters.,” *Acta Oncol. (Madr).*, vol. 49, no. 7, pp. 1012–6, 2010.

[15] F. J. Brooks and P. W. Grigsby, “The Effect of Small Tumor Volumes on Studies of Intratumoral Heterogeneity of Tracer Uptake,” *J. Nucl. Med.*, vol. 55, no. 1, pp. 37–42, 2014.

[16] R. T. H. Leijenaar *et al.*, “The effect of SUV discretization in quantitative FDG-PET Radiomics: the need for standardized methodology in tumor texture analysis.,” *Sci. Rep.*, vol. 5, no. August, p. 11075, 2015.

[17] H. J. Aerts *et al.*, “Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach,” *Nat Commun*, vol. 5, p. 4006, 2014.

[18] M. Sollini, L. Cozzi, L. Antunovic, A. Chiti, and M. Kirienko, “PET Radiomics in NSCLC: state of the art and a proposal for harmonization of methodology,” *Sci. Rep.*, vol. 7, no. 1, p. 358, 2017.

[19] M. Hatt, F. Tixier, L. Pierce, P. E. Kinahan, C. C. Rest, and D. Visvikis, “Characterization of {PET/CT} images using texture analysis: the past, the present… any future?,” vol. 44, no. 1, pp. 151–165, 2016.

[20] M. Hatt *et al.*, “18F-FDG PET Uptake Characterization Through Texture Analysis: Investigating the Complementary Nature of Heterogeneity and Functional Tumor Volume in a Multi–Cancer Site Patient Cohort,” *J. Nucl. Med.*, vol. 56, no. 1, pp. 38–44, 2015.

[21] J. Yan *et al.*, “Impact of Image Reconstruction Settings on Texture Features in 18F-FDG PET.,” *J. Nucl. Med.*, vol. 56, no. 11, pp. 1667–73, 2015.

[22] R. T. H. Leijenaar *et al.*, “Stability of FDG-PET Radiomics features: an integrated analysis of test-retest and inter-observer variability.,” *Acta Oncol. (Madr).*, vol. 52, no. 7, pp. 1391–7, 2013.

[23] M. Hatt, C. Cheze-Le Rest, A. Van Baardwijk, P. Lambin, O. Pradier, and D. Visvikis, “Impact of Tumor Size and Tracer Uptake Heterogeneity in 18 F-FDG PET and CT Non–Small Cell Lung Cancer Tumor Delineation,” *J Nucl Med*, vol. 52, no. 11, pp. 1690–1697, 2011.

[24] M. Hatt, F. Tixier, C. Cheze Le Rest, O. Pradier, and D. Visvikis, “Robustness of intratumour 18F-FDG PET uptake heterogeneity quantification for therapy

response prediction in oesophageal carcinoma," *Eur. J. Nucl. Med. Mol. Imaging*, vol. 40, no. 11, pp. 1662–1671, 2013.

[25] F. H. P. van Velden *et al.*, "Repeatability of Radiomic Features in Non-Small-Cell Lung Cancer [18F]FDG-PET/CT Studies: Impact of Reconstruction and Delineation," *Mol. Imaging Biol.*, vol. 18, no. 5, pp. 788–795, 2016.

[26] L. Azzari and A. Foi, "Variance stabilization in Poisson image deblurring," *Proc. - Int. Symp. Biomed. Imaging*, no. 4, pp. 728–731, 2017.

[27] S. W. Hasinoff, "Photon, Poisson Noise," *Comput. Vis. A Ref. Guid.*, pp. 608–610, 2014.

[28] "Alessandro Foi ICIP 2014 Tutorial T7 : Signal-Dependent Noise and Stabilization of Variance," no. Icip, 2014.

[29] F. P. Noise, "Topic 5 : Noise in Images," no. August, 2007.

[30] S. Pyatykh and J. Hesser, "MMSE Estimation for Poisson Noise Removal in Images," *arXiv1512.00717* , pp. 1–4, 2015.

[31] T. Le, R. Chartrand, and T. J. Asaki, "Denoising images with Poisson noise statistics," *Math. Model. Anal.*, vol. 27, no. 3, pp. 1–10, 2005.

[32] M. Hatt, F. Tixier, L. Pierce, P. E. Kinahan, C. C. Rest, and D. Visvikis, "Characterization of {PET/CT} images using texture analysis: the past, the present… any future?," *Eur. J. Nucl. Med. Mol. Imaging*, pp. 151–165, 2016.

[33] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst. Man. Cybern.*, vol. 9, no. 1, pp. 62–66, 1979.

[34] T. F. Chan and L. A. Vese, "Active contours without edges," *IEEE Trans. Image Process.*, vol. 10, no. 2, pp. 266–277, 2001.

[35] U. Bagci, J. Yao, J. Caban, E. Turkbey, O. Aras, and D. J. Mollura, "A graph-theoretic approach for segmentation of PET images," *2011 Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, pp. 8479–8482, 2011.

[36] W. Ju, D. Xiang, B. Zhang, L. Wang, I. Kopriva, and X. Chen, "Random Walk and Graph Cut for Co-Segmentation of Lung Tumor on PET-CT Images," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5854–5867, 2015.

[37] A. Depeursinge, A. Foncubierta-Rodriguez, D. Van De Ville, and H. Müller, "Three-dimensional solid texture analysis in biomedical imaging: Review and opportunities," *Med. Image Anal.*, vol. 18, no. 1, pp. 176–196, 2014.

[38] C. Parmar, P. Grossmann, J. Bussink, P. Lambin, and H. J. W. L. Aerts, "Machine Learning methods for Quantitative Radiomic Biomarkers," *Sci. Rep.*, vol. 5, p.

13087, 2015.

[39] C. Malsburg, “Self-organization of orientation sensitive cells in the striate cortex,” *Kybernetik*, vol. 14, no. 2, pp. 85–100, 1973.

[40] C. Cortes and V. Vapnik, “Support-Vector Networks,” *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[41] C. M. Bishop and N. Nasrabadi, “Pattern Recognition and Machine Learning,” *Pattern Recognit.*, vol. 4, no. 4, p. 738, 2006.

[42] L. Breiman, “Random Forests,” *Mach. Learn.*, vol. 45, no. 5, pp. 1–35, 1999.

[43] N. Friedman, D. Geiger, and M. Goldszmidt, “Bayesian Network Classifiers,” *Mach. Learn.*, vol. 29, pp. 131–163, 1997.

[44] Y. Freund and R. R. E. Schapire, “Experiments with a New Boosting Algorithm,” *Int. Conf. Mach. Learn.*, pp. 148–156, 1996.

[45] C. Chang and C. Lin, “LIBSVM : A Library for Support Vector Machines,” *ACM Trans. Intell. Syst. Technol.*, vol. 2, pp. 1–39, 2013.

[46] M. H. Hesamian, W. Jia, X. He, and P. Kennedy, “Deep Learning Techniques for Medical Image Segmentation: Achievements and Challenges,” *J. Digit. Imaging*, vol. 32, no. 4, pp. 582–596, 2019.

[47] J. Long, E. Shelhamer, and T. Darrell, “Long_Shelhamer_Fcn.”

[48] E. Shelhamer, J. Long, and T. Darrell, “Fully Convolutional Networks for Semantic Segmentation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, 2017.

[49] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 9351, pp. 234–241, 2015.

[50] F. Milletari, N. Navab, and S. A. Ahmadi, “V-Net: Fully convolutional neural networks for volumetric medical image segmentation,” *Proc. - 2016 4th Int. Conf. 3D Vision, 3DV 2016*, pp. 565–571, 2016.

[51] S. Asgari Taghanaki, K. Abhishek, J. P. Cohen, J. Cohen-Adad, and G. Hamarneh, “Deep semantic segmentation of natural and medical images: a review,” *Artif. Intell. Rev.*, vol. 54, no. 1, pp. 137–178, 2021.

[52] F. Yousefirizi, A. K. Jha, J. Brosch-Lenz, B. Saboury, and A. Rahmim, “Toward High-Throughput Artificial Intelligence-Based Segmentation in Oncological PET Imaging,” *PET Clin.*, vol. 16, no. 4, pp. 577–596, 2021.

[53] Philipp Seeböck, “Deep Learning in Medical Image Analysis,” vol. 2015, no.

March, pp. 221–248, 2015.

[54] T. C. Kwee, G. Cheng, M. G. E. H. Lam, S. Basu, and A. Alavi, “SUVmax of 2.5 should not be embraced as a magic threshold for separating benign from malignant lesions,” *Eur. J. Nucl. Med. Mol. Imaging*, vol. 40, no. 10, pp. 1475–1477, 2013.

[55] R. Boellaard *et al.*, “FDG PET/CT: EANM procedure guidelines for tumour imaging: version 2.0,” *Eur. J. Nucl. Med. Mol. Imaging*, vol. 42, no. 2, pp. 328–354, 2015.

[56] W. D. Travis *et al.*, “The 2015 World Health Organization Classification of Lung Tumors,” *J. Thorac. Oncol.*, vol. 10, no. 9, pp. 1243–1260, 2015.

[57] S. Rizzo *et al.*, “Radiomics: the facts and the challenges of image analysis,” *Eur. Radiol. Exp.*, vol. 2, no. 1, p. 36, 2018.

[58] F. Orlhac, M. Soussan, K. Chouahnia, E. Martinod, and I. Buvat, “18F-FDG PET-derived textural indices reflect tissue-specific uptake pattern in non-small cell lung cancer,” *PLoS One*, vol. 10, no. 12, pp. 1–16, 2015.

[59] P.-P. Ypsilantis *et al.*, “Predicting Response to Neoadjuvant Chemotherapy with PET Imaging Using Convolutional Neural Networks.,” *PLoS One*, vol. 10, no. 9, p. e0137036, 2015.

[60] S. H. Hawkins *et al.*, “Predicting Outcomes of Nonsmall Cell Lung Cancer Using CT Image Features,” *IEEE Access*, vol. 2, pp. 1418–1426, 2014.

[61] H. J. W. L. Aerts *et al.*, “Defining a Radiomic Response Phenotype: A Pilot Study using targeted therapy in NSCLC,” *Sci. Rep.*, vol. 6, no. September, p. 33860, 2016.

[62] K.-H. Yu *et al.*, “Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features.,” *Nat. Commun.*, vol. 7, p. 12474, 2016.

[63] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, “Machine learning applications in cancer prognosis and prediction,” *Comput. Struct. Biotechnol. J.*, vol. 13, pp. 8–17, 2015.

[64] M. P. Menden *et al.*, “Machine Learning Prediction of Cancer Cell Sensitivity to Drugs Based on Genomic and Chemical Properties,” *PLoS One*, vol. 8, no. 4, 2013.

[65] A. C. Haury, P. Gestraud, and J. P. Vert, “The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures,” *PLoS One*, vol. 6, no. 12, pp. 1–12, 2011.

[66] S. Ha *et al.*, “Autoclustering of Non-small Cell Lung Carcinoma Subtypes on 18F-

FDG PET Using Texture Analysis: A Preliminary Result," *Nucl. Med. Mol. Imaging (2010).*, vol. 48, no. 4, pp. 278–286, 2014.

[67] M. Vallières, C. R. Freeman, S. R. Skamene, and I. El Naqa, "A radiomics model from joint FDG-PET and MRI texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities," *Phys. Med. Biol.*, vol. 60, no. 14, pp. 5471–5496, 2015.

[68] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten, "The WEKA data mining software: An update," *SIGKDD Explor.*, vol. 11, no. 1, pp. 10–18, 2009.

[69] S. W. Hasinoff, "Photon, Poisson Noise," in *Computer Vision*, Springer US, 2014, pp. 608–610.

[70] J. Ross, Q. Morgan, and K. Publishers, "Book Review : C4 . 5 : Programs for Machine Learning," vol. 240, pp. 235–240, 1994.

[71] D. Broomhead, D. S. and Lowe, "Multivariable Functional Interpolation and Adaptive Networks," *Complex Syst.*, vol. 2, pp. 321–355, 1988.

[72] D. H. Wolpert, "Stacked Generalization," vol. 87545, no. 505, pp. 241–259, 1992.

[73] H. Sung *et al.*, "Global Cancer Statistics 2020 : GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries," vol. 0, no. 0, pp. 1–41, 2021.

[74] B. S. Greenspan, "Role of PET / CT for precision medicine in lung cancer : perspective of the Society of Nuclear Medicine and Molecular Imaging," vol. 6, no. 6, pp. 617–620, 2017.

[75] B. Foster, U. Bagci, A. Mansoor, Z. Xu, and D. J. Mollura, "A review on segmentation of positron emission tomography images," *Comput. Biol. Med.*, vol. 50, pp. 76–96, 2014.

[76] H. Zaidi and I. El Naqa, "PET-guided delineation of radiation therapy treatment volumes: A survey of image segmentation techniques," *Eur. J. Nucl. Med. Mol. Imaging*, vol. 37, no. 11, pp. 2165–2187, 2010.

[77] M. Hatt *et al.*, "The first MICCAI challenge on PET tumor segmentation," *Med. Image Anal.*, vol. 44, pp. 177–195, 2018.

[78] B. Foster, U. Bagci, A. Mansoor, Z. Xu, and D. J. Mollura, "A review on segmentation of positron emission tomography images," *Comput. Biol. Med.*, vol. 50, pp. 76–96, 2014.

[79] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-

net: Learning dense volumetric segmentation from sparse annotation," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 9901 LNCS, pp. 424–432, 2016.

[80] F. Milletari, N. Navab, and S. A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," *Proc. - 2016 4th Int. Conf. 3D Vision, 3DV 2016*, pp. 565–571, 2016.

[81] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, no. December 2012, pp. 60–88, 2017.

[82] L. Chen, G. Papandreou, and I Kokkinos, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *Ieeexplore.Ieee.Org*, vol. 40, no. 4, pp. 834–848, 2018.

[83] N. Abraham and N. M. Khan, "A Novel Focal Tversky loss function with improved Attention U-Net for lesion segmentation," 2018.

[84] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4510–4520.

[85] O. Ayyildiz *et al.*, "Lung cancer subtype differentiation from positron emission tomography images," *Turkish J. Electr. Eng. Comput. Sci.*, vol. 28, no. 1, pp. 262–274, 2020.

[86] F. Renard, S. Guedria, N. De Palma, and N. Vuillerme, "Variability and reproducibility in deep learning for medical image segmentation," *Sci. Rep.*, vol. 10, no. 1, pp. 1–16, 2020.

[87] G. Wang *et al.*, "Interactive Medical Image Segmentation Using Deep Learning with Image-Specific Fine Tuning," *IEEE Trans. Med. Imaging*, vol. 37, no. 7, pp. 1562–1573, 2018.

[88] Z. Zhong *et al.*, "Simultaneous cosegmentation of tumors in PET-CT images using deep fully convolutional networks," *Med. Phys.*, vol. 46, no. 2, pp. 619–633, 2019.

[89] L. Li, X. Zhao, W. Lu, and S. Tan, "Deep learning for variational multimodality tumor segmentation in PET/CT," *Neurocomputing*, vol. 392, no. 7, pp. 277–295, 2019.

[90] X. Zhao, L. Li, W. Lu, and S. Tan, "Tumor co-segmentation in PET/CT using multi-modality fully convolutional neural network," *Phys. Med. Biol.*, vol. 64, no. 1, pp. 1–35, 2019.

[91] A. McGuigan, P. Kelly, R. C. Turkington, C. Jones, H. G. Coleman, and R. S. McCain, “Pancreatic cancer: A review of clinical diagnosis, epidemiology, treatment and outcomes,” *World J. Gastroenterol.*, vol. 24, no. 43, pp. 4846–4861, 2018.

[92] C. Feig, A. Gopinathan, A. Neesse, D. S. Chan, N. Cook, and D. a Tuveson, “Europe PMC Funders Group The pancreas cancer microenvironment,” *Clin Cancer Res.*, vol. 18, no. 16, pp. 4266–4276, 2013.

[93] E. H. Dibble, D. Karantanis, G. Mercier, P. J. Peller, L. A. Kachnic, and R. M. Subramaniam, “PET/CT of cancer patients: Part 1, pancreatic neoplasms,” *Am. J. Roentgenol.*, vol. 199, no. 5, pp. 952–967, 2012.

[94] Z. Wang, J. Q. Chen, J. L. Liu, X. G. Qin, and Y. Huang, “FDG-PET in diagnosis, staging and prognosis of pancreatic carcinoma: A meta-analysis,” *World J. Gastroenterol.*, vol. 19, no. 29, pp. 4808–4817, 2013.

[95] H. J. Choi *et al.*, “Prognostic value of 18F-Fluorodeoxyglucose positron emission tomography in patients with resectable pancreatic cancer,” *Yonsei Med. J.*, vol. 54, no. 6, pp. 1377–1383, 2013.

[96] H. X. Xu *et al.*, “Metabolic tumour burden assessed by 18F-FDG PET/CT associated with serum CA19-9 predicts pancreatic cancer outcome after resection,” *Eur. J. Nucl. Med. Mol. Imaging*, vol. 41, no. 6, pp. 1093–1102, 2014.

[97] Y. Yue *et al.*, “Identifying prognostic intratumor heterogeneity using pre- and post-radiotherapy 18F-FDG PET images for pancreatic cancer patients,” *J. Gastrointest. Oncol.*, vol. 8, no. 1, pp. 127–138, 2017.

[98] A. Adadi and M. Berrada, “Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI),” *IEEE Access*, vol. 6, pp. 52138–52160, 2018.

[99] T. P. Alekya, M. P. Venkatesh, and P. K. T.M., “Software As Medical Device,” *Int. J. Drug Regul. Aff.*, vol. 4, no. 2, pp. 10–18, 2018.

[100] US FDA and U.S. Food and Drug Administration, “Proposed Regulatory Framework for Modifications to Artificial Intelligence / Machine Learning (AI / ML) -Based Software as a Medical Device (SaMD) - Discussion Paper and Request for Feedback,” *U.S Food Drug Adm.*, pp. 1–32, 2017.

[101] M. Reyes *et al.*, “On the interpretability of artificial intelligence in radiology: Challenges and opportunities,” *Radiol. Artif. Intell.*, vol. 2, no. 3, 2020.

[102] J. Civit-Masot, F. Luna-Perejon, S. Vicente-Diaz, J. M. Rodriguez Corral, and A.

Civit, “TPU cloud-based generalized U-Net for eye fundus image segmentation,” *IEEE Access*, vol. 7, pp. 142379–142387, 2019.

CURRICULUM VITAE

2013	B.Sc., Electrical and Electronics Engineering, Bilkent University, Ankara, TURKEY
2016	M.Sc., Biomedical Engineering, Erciyes University, Kayseri, TURKEY
2016 – 2022	Doctoral Candidate, Electrical and Computer Engineering, Abdullah Gül University, Kayseri, TURKEY
2015	Tekno Girişim Fellowship, Bilim ve Sanayi Bakanlığı, Ankara, TURKEY
2013-2016	TÜBİTAK Fellowship, Project No:113E188, Kayseri, TURKEY
2013 – Present	Teaching and Research Assistant, Electrical and Electronics Engineering, Abdullah Gül University, Kayseri, TURKEY

SELECTED PUBLICATIONS AND PRESENTATIONS

J1) Karacavus, S., B. Yılmaz, A. Tasdemir, Ö. Kayaaltı, E. Kaya, S. İçer, **O. Ayyıldız**, E. Vardareli, M.H. Asyalı, “Can Laws’ be a potential PET image texture analysis approach for evaluation of tumor heterogeneity and histopathological characteristics in NSCLC?” J Digit Imaging, 31:210 (2018)

J2) Ayyıldız, O., Z. Aydın, B. Yılmaz, S. Karaçavuş, K. Şenkaya, S. İçer, A. Erdem Taşdemir, E. Kaya, “Lung cancer subtype differentiation from Positron Emission Tomography Images” Turk J Elec Eng & Comp Sci, 28: 262 – 274 (2020)

J3) Bıçakcı, M., **O. Ayyıldız**, Z. Aydın, A. Baştürk, S. Karaçavuş, B. Yılmaz, “Metabolic Imaging based Sub-Classification of Lung Cancer”, IEEE Access, 8, 218470-218476, doi: 10.1109/ACCESS.2020.3040155 (2020)

C1) Torun, N., S. Karacavus, **O. Ayyildiz**, F. Kayaselçuk, G. N. Nursal, B. Yilmaz, M. Reyhan, A. F. Yapar, G. Zararsiz, “Prognostic value of FDG PET metabolic parameters and textural features in pancreatic adenocarcinoma”, 12th Asia Oceania Congress of Nuclear Medicine and Biology, Yokohama, Japan, 2017.

C2) Karaçavuş, S., B. Yılmaz, Ö. Kayaaltı, A. Taşdemir, E. Kaya, S. İçer, **O. Ayyıldız**, K. Eset, E. Vardareli, M.H. Asyalı, “Prognostic significance of the texture features determined using three dimensional 18F-FDG PET images: new potential biomarkers”, Annual Congress of the Society of Nuclear Medicine and Molecular Imaging, San Diego, CA, USA, 2016.