**班 级 Class**

**学 号 ID No.**

****

**XIDIAN UNIVERSITY**

本科毕业设计论文

**UNDERGRADUATE GRADUATION DISSERTATION**



|  |  |
| --- | --- |
| **题 目Topic** |  |
|  |  |
| **学 院School** |  |
| **专 业Major** |  |
| **学生姓名Name** |  |
| **导师姓名Tutor** |  |

Contents

[Abstract 1](#_Toc42020452)

[Chapter 1 1](#_Toc42020453)

[Introduction 1](#_Toc42020454)

[**1.1** **Head and neck cancer** 1](#_Toc42020455)

[**1.1.1**  **Pathological diagnosis** 2](#_Toc42020456)

[**1.1.2 Staging** 2](#_Toc42020457)

[**1.2 Computed tomography scan (CT)** 2](#_Toc42020458)

[**1.3 DICOM STANDARD** 4](#_Toc42020459)

[Chapter 2 5](#_Toc42020460)

[Materials and Methods 5](#_Toc42020461)

[**2.1 Dataset and environment settings** 5](#_Toc42020462)

[**2.2 Data Preprocessing** 6](#_Toc42020463)

[**2.3 Segmentation** 6](#_Toc42020464)

[**2.4 Feature extraction** 9](#_Toc42020465)

[**2.4.1 The** **first order (Intensity of Histogram)** 11](#_Toc42020466)

[**2.4.2 Second Order (GLCM)** 13](#_Toc42020467)

[**2.4.3 Moments** 15](#_Toc42020468)

[**2.4.4** **Region Properties** 18](#_Toc42020469)

[**2.4.5 2D Convolutions** 19](#_Toc42020470)

[Chapter 3 21](#_Toc42020471)

[Radiomics Workflow 21](#_Toc42020472)

[**3.1 Radiomics Workflow** 21](#_Toc42020473)

[**3.2 Radiomics Limitations** 23](#_Toc42020474)

[**3.2.1 Imaging Acquisition** 23](#_Toc42020475)

[**3.2.2 Image Segmentation** 23](#_Toc42020476)

[**3.2.3 Feature Extraction** 23](#_Toc42020477)

[**3.2.4 Image Processing** 24](#_Toc42020478)

[Chapter 4 25](#_Toc42020479)

[Experimental results 25](#_Toc42020480)

[Chapter 5 26](#_Toc42020481)

[Conclusions And Future Direction 26](#_Toc42020482)

[**5.1 Conclusions** 26](#_Toc42020483)

[**5.2 Future Direction** 26](#_Toc42020484)

[**5.2.1 Feature Selection** 26](#_Toc42020485)

[**5.2.2 Machine learning algorithms** 27](#_Toc42020486)

[**References** 28](#_Toc42020487)

[**ACKNOWLEDGEMENTS** 32](#_Toc42020488)

[Appendix A 33](#_Toc42020489)

[Appendix B 38](#_Toc42020490)

# Abstract

In the past decade, the sector of medical image analysis has mature exponentially including head and neck cancer, with a rised range of pattern recognition tools and an increased size of data set. These advances have expedited the event of processes for high-throughput extraction of quantitative options that lead to the conversion of images into mineable information and also for the later analysis of those information for decision support; this practice is known as radiomics. This is generally in contrast to the standard practice of treating medical data as images supposed exclusively for visual interpretation. Radiomic information consist first order, second order, and higher order statistics. These information are combined with different patient data and are well-mined with refined bioinformatics tools to develop models which will probably improve diagnostic, prognostic, and predictive accuracy. as a result of radiomics analyses are meant to be conducted with standard of care image datasets, it's conceivable that conversion of digital pictures to mineable information will eventually become routine practices. Using Radiomics and texture analysis we are able to extract a large number of quantitative features, analyzing their properties to include and help in clinical decision-making. Head and neck cancers present a unique set of diagnostic and therapeutic challenges naturally of its complicated anatomy and heterogeneousness. To predict the likely course of Head and Neck cancer biomarker derived from routine, standard of care, imaging information, and providing support throughout the follow of the patient, in some cases Radiomics and texture analysis will facilitate avoiding the requirement for biopsies. Here we will help to understand the outline of radiomic texture analysis strategies, review the developments of radiomics in head and neck cancer applications, discuss unmet challenges, and can make a case for basic ideas concerning radiomics and discuss recent progress and results associated with Head and Neck cancer.

**Keywords**: Head and neck cancer, Computed tomography, Dicom, feature extraction, Machine learning

# 

# 

# Chapter 1

# Introduction

Head and neck cancer represent the sixth most common malignancy worldwide, with around 800,000 new cases and 320,000 deaths in 2015 [1]. Worldwide, more than 500,000 patients are diagnosed with head and neck squamous cell carcinoma (HNSCC) each year and thus posing substantial economic burdens Head and neck cancer has high incidence in southern China especially in Hong Kong. However medical imaging has helped in the diagnosis and treatment of Head and neck cancer [2]. Head and neck cancers comprise a heterogeneous group of tumors arising from the oral cavity, pharynx, larynx, sinuses, and salivary and thyroid glands. Surgery is still one of the major treatment options for this disease, and with the primary aim of maximizing tumor removal while minimizing the damage to healthy tissue. Surgery may cause unnecessary removal of normal tissues which can leave patients with functional and aesthetic problems, such as chewing, swallowing, or speaking. And there are lots of chances that a diseased tissue is not completely removed, because of that cancer likely to recur or may cause another surgery. A positive surgical margin is associated with a poor prognosis in terms of increased local recurrence of cancer and decreased overall patient survival. Therefore, the tumor tissues are needed to be highly accurate for maximizing the efficiency of the surgical treatment and increase the chances of patient’s subsequent quality of life, both of which help in the cost savings.

## **1.1** **Head and neck cancer**

Head and neck cancer incidence rate ranks sixth in males and the mortality rate ranks seventh in China ﻿[3]. The most common pathological type is squamous cell carcinoma. With the exception of nasopharyngeal carcinoma, which is mainly caused by Epstein-Barr virus, most other head and neck squamous cell carcinomas are due to cigarette smoking and alcohol abuse ﻿[4]. In recent years, the incidence of oropharyngeal cancer has significantly increased in Europe and the United States, mainly due to human papillomavirus (HPV) infection, but its exact infection rate is still unclear in China ﻿[5].  
The diagnosis of primary head and neck cancer mainly depends on oral or endoscopic mass biopsy, whereas lymph node puncture or biopsy is helpful in cancer staging. Other countries usually advocate upper gastrointestinal endoscopy (panendoscopy) under general anesthesia for biopsies on suspicious lesions, which helps to increase the accuracy of diagnosis and the possibility of identifying a second primary cancer.

### **1.1.1** **Pathological diagnosis**

The pathology of head and neck cancer is critical for identifying stages and deciding treatment strategies [6]. For both biopsy and punctured specimens, it is essential to distinguish the benign lesion from the malignant tumor, and determine its histological type. If necessary, immunohistochemical staining should be applied. For surgical specimens obtained from radical resection of head and neck squamous cell carcinoma, information including tumor size, differentiation, margin, vascular invasion, peripheral nerve infiltration, bone or cartilage infiltration, site, number of lymph node metastases, and extracapsular invasion is required. For oral cancer, it is necessary to clarify the depth of tumor invasion, which is helpful in guiding the follow-up treatment strategy [7]. For oropharyngeal cancer, immunohistochemical examination of p16 can be performed if possible, to determine whether it is associated with HPV infection, although current guidelines do not recommend that individualized treatment strategy be based on this test result[7].

### **1.1.2 Staging**

Current guidelines apply the Union for International Cancer Control/American Joint Committee on Cancer (AJCC) Tumor-Node-Metastasis staging system (7th edition), due to its popularity. The 8th edition was published on January 1, 2018, but we did not use it here, as it recommends an HPV test in patients with oralpharynx cancer and requires the detection of the depth of invasion in patients with oral cancer, which are not routine tests performed in China﻿ [8].

## **1.2 Computed tomography scan (CT)**

Enhanced computed tomography (CT) or magnetic resonance imaging (MRI) scans of the primary lesion are common methods for diagnosing head and neck cancer, both of which have advantages and disadvantages. CT has the advantages of simplicity, rapidity, and wide availability. On the other hand, it has the disadvantage of radiation exposure, which is not appropriate for patients with iodine allergy or severe renal insufficiency. MRI can show soft tissue with higher resolution than CT and provides a variety of imaging measurements. It is especially useful for tumors originating from the oral cavity, oropharynx, and nasopharynx, and has excellent resolution for the skull base and nerves. While, the disadvantage of MRI is that it is time-consuming and relatively expensive, and is not suitable for patients with metal implants or with claustrophobia. In addition, for the laryngeal and hypopharyngeal organs, it is easy to cause artifacts due to involuntary swallowing during MRI examination. The neck is the most common area for lymph node metastasis from head and neck cancer. Neck-enhanced CT is a standard staging method, especially during the characterization of typical features of lymph node necrosis. The lung is the most common site for distant metastasis from head and neck cancer. Chest CT is the standard staging method that also helps assess other pulmonary diseases such as chronic bronchitis.

Positron emission tomography (PET)/CT mainly uses 18F-FDG as the tracer, and has been extensively investigated for the management of head and neck cancer in recent years [9]. For primary lesions, because PET/CT is commonly performed with low-dose plain CT, its resolution is not as good as that in enhanced CT and can result in false positive or false negative results. For cervical lymph node and distant metastases, some meta-analyses have shown that PET/CT had certain advantages﻿ [10]. A prospective study showed that the combination of PET/CT with conventional staging methods changed the treatment strategy in 13.7% of patients [11]. Currently, the National Comprehensive Cancer Network (NCCN) recommends PET/CT for pre-treatment examination in patients with stage III/IV cancer﻿ [12]﻿.

Medical imaging could be a speedily developing branch of recent medication. it's within the past few decades evolved into a extremely subtle diagnostic tool. it's improved the study of human internal anatomy and to an extent physiology and detection of pathologies that were antecedently not possible. At this stage of its development, detection of lesions and their interpretation is turning into an automatic computer-aided method. It indicates that machine vision has become an emerging part of radiology and imaging in medicine. this can be as a results of advances in medical imaging technology and engineering that have greatly increased the interpretation of medical pictures and contributed to early identification. With the more use of Computed tomography (CT) the use of computers in calculating and facilitating their processing and analysis has become necessary.

Using CT images, an attempt has been made to improve the quantitative and qualitative similarity measures of accuracy for segmentation and classification of abnormal tumor regions. The different types of abnormalities of the CT images caused by the tumors are categorized as benign and malignant tumors. benefits of mistreatment CT embrace sensible detection of calcification, hemorrhage and bony detailbenefits of using CT embrace sensible detection of calcification, hemorrhage and bony detail. It also advantage of lower cost, short imaging times and widespread availability. computed tomography (CT) scanners are more widely available in the communities and may be accessed much more easily [13]. computerized tomography examinations don't seem to be solely cheaper than resonance imaging (MRI) however additionally quicker to perform. Thus, taking the time-critical nature of early stroke identification into thought, NCCT is that the most popular first-line imaging tool [14]. Computed tomography and other neuroimaging procedures will, however, not benefit the patient until the images have been accurately interpreted. The challenges related to the visual interpretation of stroke CT images are dearth of neuroradiologists and also the human errors of interpretation and diagnosis. Errors in visual interpretation result from poor technique, failures of perception, lack of information and misjudgements. Visual interpretation is improved upon by texture analysis which is able to build it potential for machine-controlled computer-aided approach to be used as a second opinion for clinicians, particularly in equivocal cases. Automatic technique of stroke detection follows an equivalent pattern as visual analysis and interpretation employed by radiologists.

## **1.3 DICOM STANDARD**

Significance of DICOM images in medical field is widespread. Imaging and Communication in Medicine (DICOM) is a standard for storing, handling, printing, and transmitting information between the digital imaging computer systems in medical environments. It includes a network communication protocol and file format definition. The communication protocol is an application protocol that uses TCP/IP to communicate between different systems. DICOM files can be exchange between two entities that are capable for receiving image and patient data in DICOM format. The National Electrical Manufacturers Association (NEMA) has the copyright to this standard. DICOM enables the integration of scanners, workstations, servers, printers, and network hardware from multiple manufacturers into a picture archiving and communication system (PACS). The different devices come with DICOM standard conformance statements which clearly state the DICOM classes they support. DICOM standard is the third version of a standard developed by American College of Radiology (ACR) and National Electrical Manufacturers Association (NEMA). DICOM is different from other data formats in that it groups information into data sets, that means a file of a chest X-Ray image, for example, actually it contains the patient ID within the file, that means the image can never be separated from this information in any condition. DICOM information object consists of variety of attributes, together with things like name, ID etc., and conjointly one special attribute containing the image pixel. A single DICOM image can only contain one attribute pixel data. This corresponds to a single image for many modalities. Pixel data can be compressed using a variety of standards, including JPEG, JPEG Lossless, Run-length encoding, and JPEG 2000 etc. The DICOM standard facilitates of medical imaging equipment by specifying

* For network communications, a set of protocols to be followed by devices claiming conformance to the DICOM standard.
* The syntax and semantics of commands and associates the data which can be exchanged using these protocols.
* For media communication, a set of media storage devices to be followed by devices claiming conformance to the DICOM standard, as well as a file format and a medical dictionary structure to facilitate access to the images and related information on interchange media.
* Information that must be supplied with an implementation for which conformance to the DICOM standard is claimed.

# Chapter 2

# Materials and Methods

## **2.1 Dataset and environment settings**

In the medical world, to ingress the real medical images like MRI, PET or CT scan and to resume a research are a very complicated because of privacy problem and heavy technical deadlocks. However these days medical image dataset is publically available on internet for study and research purposes so, I have collected the small of head and neck cancer patients dataset, which contains 28 patients. Each patient folder contains two subfolders. One folder contains the multiple DICOM standard images of that patient and another subfolder contains RTSS.mat file which contains the information about the segmentation of each image of the patient. Our objective is to extract the valuable information from the DICOM images by extracting the region of interest using segmentation which will help to predict the seriousness of the head and neck cancer. For feature extraction process here we used python. We have used various python libraries for extracting the various features from an CT image. Among these methods, Support Vector Machine (SVM), a supervised learning algorithm, has showed great superiority with small sample size of data. The DICOM standard is complex and there are number of different tools to work with DICOM files, I have used the pydicom package, which is a package that can be imported towards working with images in python.

I have used Python 3.5.0 for processing the DICOM files and RTSS.mat files so that we can extract the features. To implement the whole process I have installed different libraries such as:

* Installed Pydicom for reading the DICOM (Digital Imaging and Communications in Medicine) files.
* Installed OpenCV for manipulating the image like resizing the image.
* Installed MatPlotLib for visualization of the image files and various presentation purposes.
* Installed pandas (used for reading operations on csv files and computation of confusion matrix), os (File and directory manipulations), scikit-learn (to compute confusion matrix) ,
* Installed mat73 to read .mat file
* Installed numpy to work with numpy array
* Installed glob to get the list of directories and files inside the folder
* Installed xtract\_features to extract some features
* Installed scipy.stats to get skew and kurtosis etc.

## **2.2 Data Preprocessing**

Preprocessing stages involves to the source data analysis and information extraction, and are normally class as radiometric or geometric improvements. Radiometric improvements include correcting the data for sensor unreliable and redundant sensor or atmospheric noise, removal of nonbrain voxels and transmitting the data so they exactly describe the reflected or emitted radiation monitored by the sensor. Medical Images do not show powerful distortion, but can affected from artifacts due to various factors like surgical equipment’s, hand movement, radio frequency noise from bipolar coagulation. In this preprocessing stage is divided into two section. First section for removal of film artifacts and second one for removal of unwanted portion of MRI. We have the ROI(Region of Interest) information in RTSS.mat file which is accessible in the form of numpy Array so we also need the images in the form of numpy array for the masking process. So we converted all the images in the form of numpy array. Since the size of ROI segmented array data is same as the patient’s image numpy array data we can mask both images for further processing.

## **2.3 Segmentation**

Image segmentation is also outlined because the method of dividing an image into disjoint homogenous regions. These regions typically contain similar objects of interest or a part of them. The extent of homogeneity of the segmental regions may be measured using some image property. On the opposite hand, cluster may be outlined because the best partitioning of a given set of n data points into c subgroups, specified data points belonging to constant cluster area unit as the same as one another as doable whereas data points from 2 completely different groups share the utmost distinction [15]. The goal of segmentation is to change the illustration of a picture into completely different segments that's additional meaningful and easier to investigate. Image segmentation is often used to locate objects and boundaries in images. It is also the method of assigning a label to every pixel in an image such that pixels with the same label share certain visual characteristics. The result of image segmentation is the same size as the complete image. Image segmentation is the process of dividing an image into disjoint homogenous regions. These regions typically contain similar objects of interest. Pixel intensity is used for measuring the homogeneity of the segmented regions. There are so many practical applications of image segmentation such as Medical Imaging, Locate objects in satellite images, Face recognition, Fingerprint recognition, Traffic control systems, and Machine vision [16]. Fortunately, for our dataset segmentaion is already done and saved all segmented information in RTSS.mat file which is later extracted using mat73 library in the form of numpy array in python.

Let's review the raw image and its corresponding mask(extracted segmentation data from RTSS.mat file) which are numbered as figure 1 as One of the images from the dataset, figure 2 as Mask corresponding to figure 1 image extracted from RTSS.mat file, figure 3 as Applied Binary mask(fig.2) to the image(fig.1), figure 4 Overlayed image of mask and image for better understanding.



Fig 1. One of the images from the dataset

|

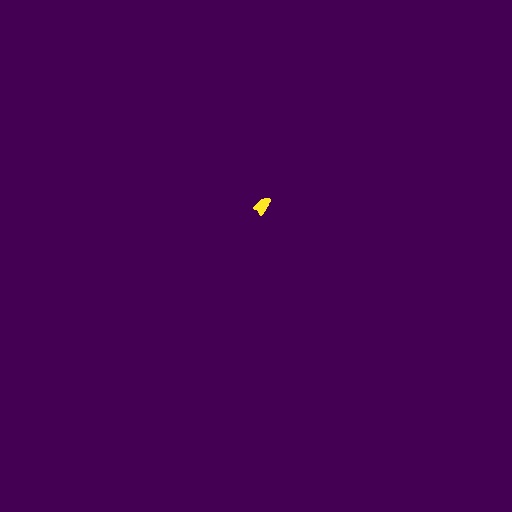


Fig 2. Mask corresponding to figure 1 image extracted from RTSS.mat file

|

|



Fig 3. Applied Binary mask(fig.2) to the image(fig.1)

|

|

|



Fig 4. Overlayed image of mask and image for better understanding

|

## **2.4 Feature extraction**

It is a method of converting input data into a set of features in a proper format. It is a process of obtaining useful and desired information from the image. After segmentation, features of lesion image extracted depending on the detail information of lesions like as color, shape, texture, or context.

Feature extraction is used to simplifying the amount of resources needed to describe a large amount of data accurately and efficiently. Number of variables large in Analysis then needed a large amount of memory and computation power. Feature extraction is a method of constructing combinations of the variables and it provide accuracy and efficiency for the data.

In this paper, we used statistical texture-based feature extraction. Two types of texture feature considered in this study, like first-order and second-order statistics. Consider intensity histogram in first-order and GLCM in second-order statistics. In this approach, features are extracted from the lesion using first order (intensity of histogram) and second-order (GLCM).

Our aim in this research paper is to extract the various features from the head and neck cancer patient’s dataset and combine them in a single file for futher processing. Firstly we take one patient folder images at a time and convert all the dicom data to numpy array file and save it as npy file for further use. Then we extracted the segmentation from RTSS.mat file which are mask and mask are binary array of same size as the patient’s image. Since all images does not have the segmentation and applying mask over each image in this way may took time and reduce the image quality so we keep the segmentation which has ROI. As we extracted the ROI data we can combine patient image and segmentation to get masked image, since the patient’s image and segmentation has the same size we can mask them easily without reshaping them. I am saving the images in different folders like masks, masked and ovelayed images for futher processing, Instead of saving images in folders we can create a npy file for them which could be a lot easier to maintain however just to visualize the difference I saved all of them in form of image. As we get the ovelayed image, we convert it into Grayscale image using Cv2 library in python and we can start the process of extracting various features which are already explained above.

Some of features are extracted manually and some are extracted using open source libraries. I have used for loop to iterate through all the patients and creating a csv file containing all features for each patient in a folder. As we are able to create the csv file for each patient we combine all csv file in one using pandas library in python. In this manner we will be having 28 different csv files which at the end combined as one csv file for further use.

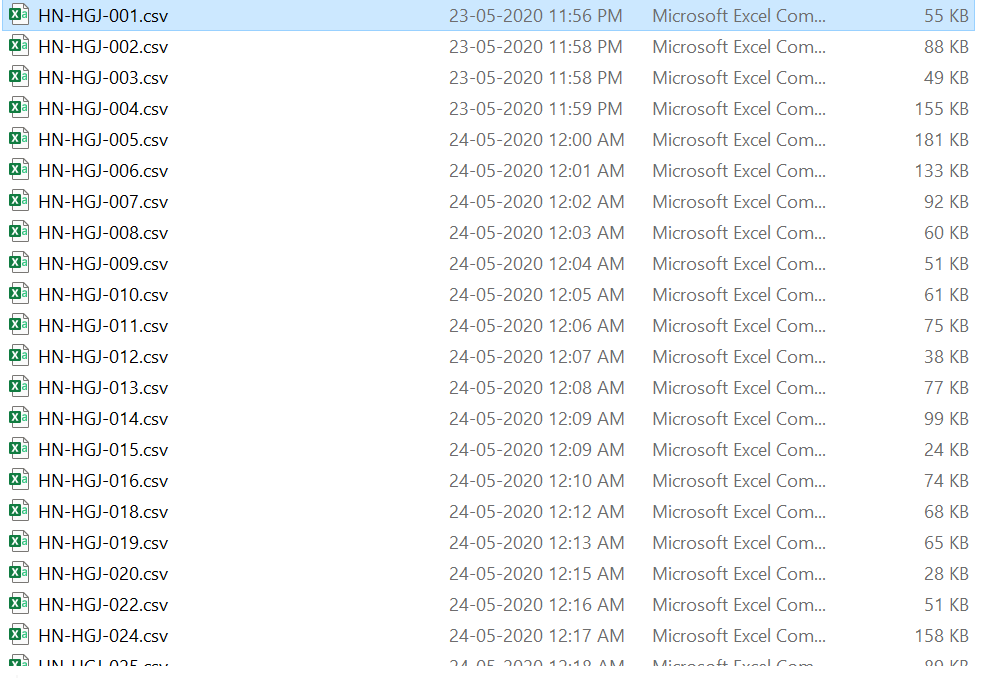


Fig 5. Extracted features for every patient and saved in csv file



Fig 6. Conbined all csv file into one file named as final.csv

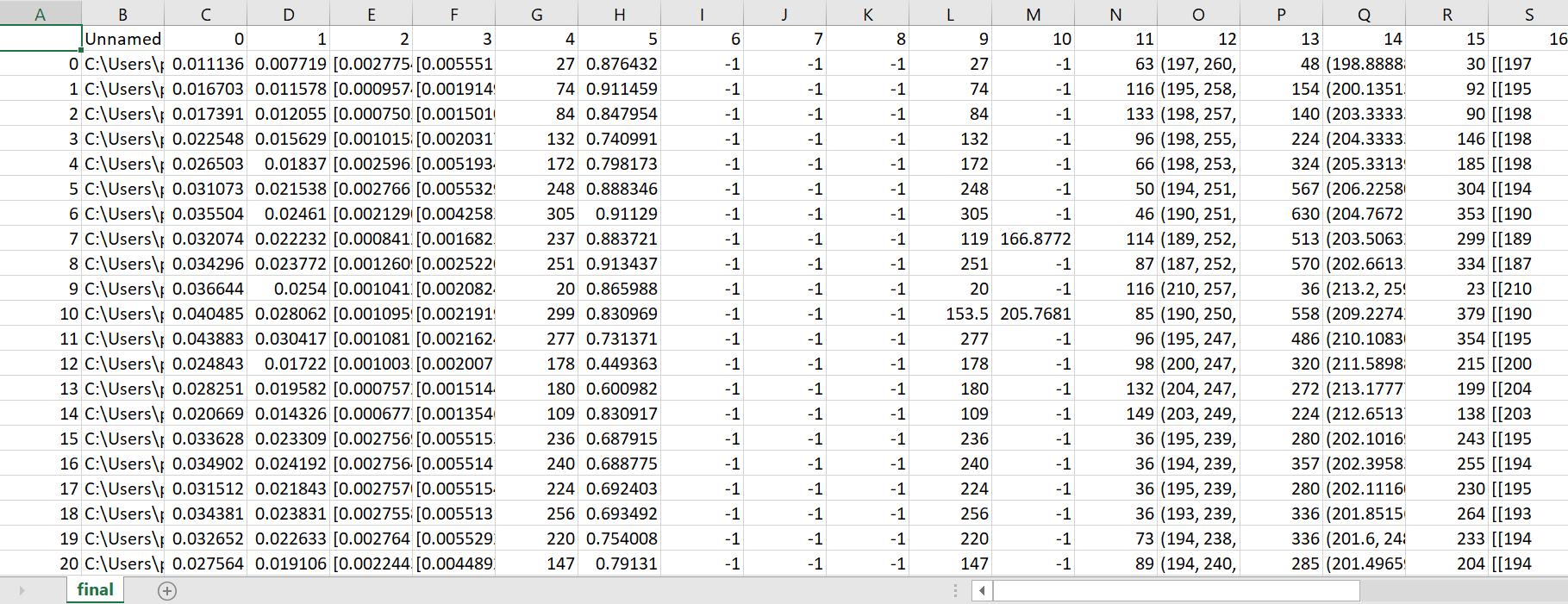


fig 7. Some part of features from final.csv

### **2.4.1 The** **first order (Intensity of Histogram)**

It provides various statistical features of the intensity of histogram of an image. It depends on individual pixel values and does not relate to neighboring values. The first- order statistics are intensity histogram and intensity features. Histogram is the probability of occurrence of a pixel of particular grey level (intensity) [17]. It does not consider the spatial relationships, and correlations, between pixels. The main advantage of the histogram is its simplicity by the use of standard descriptors such as mean and variance to characterize texture data. It included mainly properties like as mean, varience, skewness, kurtosis, Entropy.

First-order statistics or Intensity of Histogram describe the distribution of voxel intensities within the image region defined by the mask through commonly used and basic metrics[18].

Let :

* X be a set of Np voxels included in the ROI.
* P(i) be the first order histogram with Ng discrete intensity levels, where Ng is the number of non-zero bins, equally spaced from 0 with a width.
* P(i) be the normalized first order histogram and equal to P(i)/Np

1. **Mean**

The average gray level intensity within the ROI.

(2-1)

1. **Variance**

Variance is the the mean of the squared distances of each intensity value from the Mean value. This is a measure of the spread of the distribution about the mean. By definition, variance=﻿

(2-2)

1. **Skewness**

Skewness measures the asymmetry of the distribution of values about the Mean value. Depending on where the tail is elongated and the mass of the distribution is concentrated, this value can be positive or negative. Skewness is the measure of the asymmetry of a histogram (frequency distribution). A histogram with normal distribution is symmetrical. In other words, the same amount of data falls on both sides of the mean. Skewness for a normal distribution will be 0 [20]. The direction of skewness is “to the tail.” The larger the number, the longer the tail. If skewness is positive, the tail on the right side of the distribution will be longer. If skewness is negative, the tail on the left side will be longer.

(2-3)

Where μ3 is the 3rd central moment.

1. **Kurtosis**

Kurtosis could be a measure of the ‘peakedness’ of the distribution of values within the image ROI [19]. A higher kurtosis implies that the mass of the distribution is targeted towards the tail(s) instead of towards the mean[18]. A lower kurtosis implies the reverse: that the mass of the distribution is concentrated towards a spike near the Mean value. Kurtosis is a measure of the combined weight of the tails in relation to the rest of the distribution. As the tails of a distribution become heavier, the kurtosis value will increase. As the tails become lighter the kurtosis value will decrease. A histogram with a normal distribution has a kurtosis of 0. If the distribution is peaked (tall and skinny), it will have a kurtosis greater than 0 and is said to be leptokurtic. If the distribution is flat, it will have a kurtosis value less than zero and is said to be platykurtic.

(2-4)

Where μ4 is the 4th central moment.

1. **Entropy**

Entropy specifies the uncertainty/randomness in the image values. It measures the average amount of information required to encode the image values.

(2-5)

Here, ϵ is an arbitrarily small positive number (≈2.2×10−16≈2.2×10−16).

1. **Shannon entropy**

(2-6)

Where N is the number of gray levels (256 for 8-bit images), is the probability of a pixel having gray level, and is the base of the logarithm function.

### **2.4.2 Second Order (GLCM)**

The co-occurrence matrix is also known as second-order histogram that analyses the grey-level distribution of pairs of pixels. In grey-level co-occurrence matrix method, the probability of finding a pixel with a defined grey level at a defined distance and a defined angle from another pixel with defined grey level is calculated. So, the co-occurrences of pixel pairs are calculated in vertical, horizontal and 2 diagonal directions, as well as distances up to 5 pixels [21]. An essential feature of this arrangement is that each pixel has eight nearest neighbours connected to it except when the pixel is located at the periphery. A very simple illustration of grey-level co-occurrence matrix as relative positions of pixels of the same grey-level intensities Statistically, GLCM is a method of examining texture that considers the spatial relationship of pixels in the GLCM. The texture is characterised by the GLCM supported on how often pairs of pixels with specific values and in a specified spatial relationship occur in an image. GLCM stands for gray level co-occurrence matrix where the number of rows and columns is equal to the number of distinct gray levels. Generally, it consists of fourteen features. Out of these, only five features used in this study like Energy, Contrast, Correlation, and Homogeneity [22].

A Gray Level Co-occurrence Matrix (GLCM) of size Ng×Ng describes the second-order joint probability function of an image region constrained by the mask and is defined as P(i,j|δ,θ). The (i,j)th component of this matrix represents the quantity of times the mixture of levels i and j occur in two pixels within the image, that are separated by a distance of δ pixels along angle θ. The distance δ from the center voxel is defined as the distance according to the infinity norm. For δ=1, this results in 2 neighbors for each of 13 angles in 3D (26-connectivity) and for δ=2 a 98-connectivity (49 unique angles) [19].

Let:

* P(i,j) be the co-occurence matrix for an arbitrary δ and θ
* p(i,j) be the normalized co-occurence matrix and equal to P(i,j)/∑P(i,j)
* Ng be the number of discrete intensity levels in the image
* px(i)=﻿ P(i,j) be the marginal row probabilities
* py(i)=﻿P(i,j) be the marginal column probabilities

1. **Correlation**

Correlation reflects the consistency of image texture. It Measures the linear dependency of gray levels on those of neighboring pixels; it provides a measure similar to autocorrelation methods.

(2-1)

Correlation is a value between 0 (uncorrelated) and 1 (perfectly correlated) showing the linear dependency of grey level values to their various voxels within the GLCM [23].

in which *μx*, *μy*, *σx*, *σy* and were described as follows:

μx be the mean gray level intensity of px and defined as

(2-1.1)

μy be the mean gray level intensity of py and defined as

(2-1.2)

σx be the standard deviation of px and σy be the standard deviation of py.

1. **Homogeneity**

Measures the smoothness (homogeneity) of the gray level distribution of the image; it is (approximately) inversely correlated with contrast, if contrast is small, usually homogeneity is large.

(2-2)

1. **Contrast**

Contrast represents the amount of local gray level variation in an image and a high value of contrast may indicate the presence of edges, noise, or wrinkled textures in the image [24].

(2-3)

It is a measure of the local intensity variation, favoring values away from the diagonal (i=j) [19]. A larger value correlates with a greater disparity in intensity values among neighboring voxels. Contrast reflects the sharpness of images and the depth of texture grooves. Deeper texture grooves were related to high contrast and higher visual sharpness; on the contrary, low contrast led to shallow grooves and blurred pictures. Larger number of pixels with high contrast in grayscale is related to higher values of contrast textures within the image [25].

1. **Energy**

Energy is the square sum of each matrix element, reflects the grayscale distribution homogeneity of images and texture crudeness. Similar values of all co-occurrence matrix resulted in little energy profiles; on the contrary, high energy may be expected just in case of unequal values among co-occurrence matrix values [25].

(2-4)

Energy is a measure of homogeneous patterns in the image. A larger Energy implies that there are a lot of instances of intensity value pairs within the image that neighbor each other at higher frequencies [19].

### **2.4.3 Moments**

* 24 variant image moment values
* Hu Moments

#### **2.4.3.1** **Moments**

Image moments are helpful in describing objects after segmentation and play a necessary role in object recognition and shape analysis. Images moments might be employed for pattern recognition in images. Simple image properties are extracted via raw moments which is area or sum of grey levels. Image moments can be described as weighted average of image pixel intensities [22].

For simplicity, let us consider a single channel binary image I. The pixel intensity at location (x,y) is given by I(x,y). Note for a binary image I(x,y) can take a value of 0 or 1.

The simplest moment we can define is given below

(2-1)

In above equation we are calculating the sum of all pixel intensities. We can also say that, all pixel intensities are weighted only based on their intensity, however not based on their location in the image.

For a binary image, the above moment can be understood in a few other ways

* It is the number of white pixels in binary image (i.e. intensity = 1).
* It is area of white region in the binary image.

Let’s look at some more higher level moments.

(2-2)

In the above equation i and j are integers (e.g. 0, 1, 2 …). These moments are referred as raw moments to differentiate them from central moments [26].

the moments for the above mentioned equation depends on the intensity of pixels and their location in the image. So these moments are getting some idea of shape. Image moments capture data about the shape of a blob in a binary image since they contain information related to the intensity I(x,y), as well as position x and y of the pixels.

* **Centroid using Image Moments**

The centroid of a binary blob is its centre of mass directly. The centroid ﻿ is calculated from the following formula.

(2-3)

* **Central Moments**

Similar to the raw image moments, Central moments are also same, except that we subtract off the centroid from the x and y in the moment formula.

(2-4)

The above central moments are actually translation invariant. Let’s say, Position of the blob in image is not matters, if the shape of blob is the same, the moments will be the same. we can also make the moment invariant to scale and for that we need normalized central moments as shown below.

(2-5)

Central moments are translations invariant only, however normalized central moments are both translation and scale invariant [26].

#### **2.4.3.2 HU Moments**

Hu moment invariants or Hu moments are a set of 7 numbers which are calculated using central moments that are invariant to image transformations. The first six moments have been proved to be invariant to translation, scale, and rotation, and reflection. whereas the seventh moment’s sign changes for image reflection.

It is helpful that central moments are translation invariant. however that's not enough for shape matching. we'd prefer to calculate moments that are invariant to translation, scale, and rotation. Luckily, we can indeeed calculate such moments and that they are referred to as Hu Moments. They will be constant for a given shape, no matter how it's shifted, rotated, or scaled.Luckily, we don’t need to do all the calculations in OpenCV as we’ve got a utility function for Hu Moments. In OpenCV, we use HuMoments() function to calculate the Hu Moments of the shapes present in the input image[26].

### **2.4.4** **Region Properties**

Image regions, also known as objects, connected components, or blobs, have properties such as area, center of mass, orientation, and bounding box [27]. As we get the ROI masked image we need to collect the region properties of masked image. There are many region properties such as maximum area occupancy, mean of areas of all the regions, eccentricity of the maximum area region and eccentricity of the ellipse that has the similar second-moments as the region, returned as a scalar. The eccentricity is calculated by taking ratio of the distance between the foci of the ellipse and its major axis length. The value is between 0 and 1. (0 and 1 are degenerate cases. An ellipse whose eccentricity is 0 is actually a circle, while an ellipse whose eccentricity is 1 is a line segment.), euler number of the max area region in which Number of objects in the region minus the number of holes in those objects, returned as a scalar [28]. This property is supported only for 2-D label matrices. regionprops uses 8-connectivity to compute the Euler number (also known as the Euler characteristic), solidity of the max area region in which proportion of the pixels in the convex hull which are also within the region, returned as a scalar [29]. Solidity is Computed as Area/ConvexArea, perimeter of the max area region Distance around the boundary of the region returned as a scalar. regionprops calculates the perimeter by calculating the distance between each adjoining pair of pixels around the border of the region. If the image contains discontiguous regions, regionprops returns unexpected results, standard deviation of areas of all the segmented regions, otsu threshold of the image Otsu’s method is an adaptive thresholding way for binarization in image processing. By going through all possible threshold values (from 0 to 255), it can find the optimal threshold value of input image, coordinates of the bounding box of the max area region and area of the bounding box, Position and size of the smallest box containing the region, returned as a 1-by-(2\**Q*) vector. The first *Q* elements are the coordinates of the minimum corner of the box. The second *Q* elements are the size of the box along each dimension. For example, a 2-D bounding box with value [5.5 8.5 11 14] indicates that the (*x*,*y*) coordinate of the top-left corner of the box is (5.5, 8.5), the horizontal width of the box is 11 pixels, and the vertical height of the box is 14 pixels [29], centroid of the max area region, convex area of the max area region, all the coordinates of the max area region, equivalent diameter of the max area region, filled area of the max area of the region, inertia tensor of the max area region, eigen values of the inertia tensor of the max area regions, label of the region, local centroid of the max area, major axis length of the ellipse of the max area region is Length (in pixels) of the major axis of the ellipse that has the same normalized second central moments as the region, returned as a scalar., minor axis length of the ellipse of the max area region is Length (in pixels) of the minor axis of the ellipse that has the same normalized second central moments as the region, returned as a scalar, and the orientation which is Angle between the *x*-axis and the major axis of the ellipse that has the same second-moments as the region, returned as a scalar. The value is in degrees, ranging from -90 degrees to 90 degrees.

## **2.4.5 2D Convolutions**

For edge detection*,* sharpening and blurring the image we use 2D convolutions. kernel In image processing is a convolution matrix or masks which can be used for blurring, sharpening, embossing, edge detection and more by doing a convolution between a kernel and an image. Convolution is using a ‘kernel’ to extract certain ‘features’ from an input image. A kernel is a matrix, That is move across the image and multiplied with the input such that the output is enhanced in a certain desired manner [30].

There are 14 convolution kernels/matrices available inside the package as follows,

identity, edge-all, edge-H, edge-V, sharp, gauss-3, gauss-5, boxblur, gradient-H, gradient-V, sobel-H, sobel-V, emboss

**Table 2.1** List of all features that are being extracted

|  |  |  |  |
| --- | --- | --- | --- |
| **Histogram based**  **image intensity** | **GLCM** | **Moments** | **Geometry** |
| Mean | Correlation | 24 variant image  moment values | maximum area |
| Varience | Homogeneity | HU moments | mean of areas |
| Skewness | Contrast |  | Eccentricity |
| Kurtosis | Energy |  | Euler number |
| Entropy |  |  | Solidity |
|  |  |  | Perimeter of area |
|  |  |  | Standard deviation |
|  |  |  | otsu threshold |
|  |  |  | Coordinates of the boundary box |
|  |  |  | Area of the boundary box |
|  |  |  | Centroid |
|  |  |  | Convex area |
|  |  |  | Equivalent diameter |
|  |  |  | Filled area |
|  |  |  | Inertia tensor |
|  |  |  | Eigen values |
|  |  |  | Label of the region |
|  |  |  | Local centroid |
|  |  |  | Coordinates and Orientation |

# 

# Chapter 3

# Radiomics Workflow

## **3.1 Radiomics Workflow**

* High-quality and standardized imaging data must be collected. The region of interest (ROI), represented by the tumor, metastasis, or parts of it is manually/automatically defined, and therefore the volume of interest (VOI) is outlined.
* Collections of datasets from clinical practice can be collected to perform retrospective analysis in order to collect the basic radiomic feature extraction and statistical and predictive systems for prospective analysis.
* Definition and segmentation of the ROI can be defined as In each subject or object, capture of radiomic imaging information can be performed using a manual, semiautomatic or automatic approach.
* Radiomic feature are extracted from the tumor ROI concerning information related to image shape, intensity, and texture.
* To understand clinical application, an efficient series of iterative processes is required for reproducible and consistent extraction of radiomics data (Figure1)[31]. This “radiomics workflow” start with collection of high quality images with a standardized protocol. Segmentation of the tumor part is then performed, followed by feature extraction from the defined tumor region. The extracted features that gives the best performance, stability, or other defining metric are then selected for the further use in clinical applications﻿.

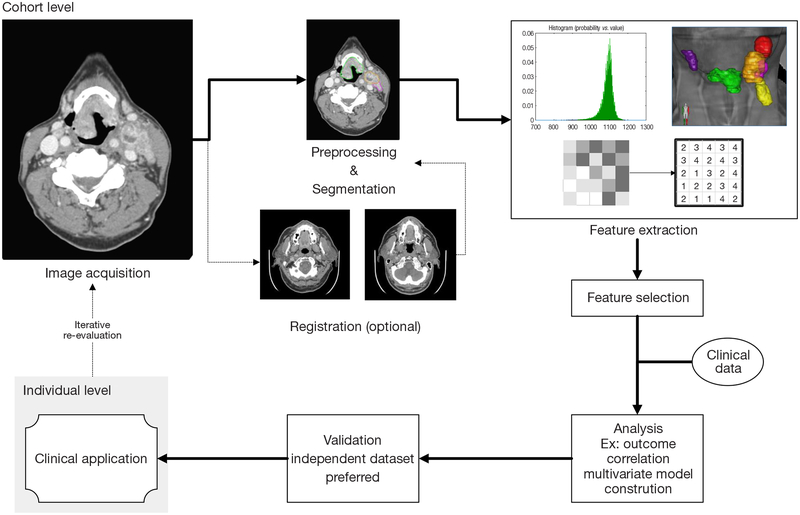


Fig 8. The radiomics workflow

|

|

The “radiomics workflow” involves a series of unvarying steps for reproducible and consistent extraction of imaging information. This process include steps such as image acquisition, tumor segmentation, feature extraction, and feature selection. The chosen features will then be analyzed for outcome correlation and potential incorporation into predictive models. In addition, validations should be done against completely independent large set of data, preferably from other institutions [32].

Methods for texture analysis in head and neck cancers are somewhat similar to those used in other organ sites. which include first order and second order texture analysis methods as well as various transformation based methods like moments. mostly “direct” features are those based simply on intensity values within a ROI and other “direct” features can be extracted from the shape of the ROI. For texture analysis from our data set, we need to extract the Segmentation from RTSS.mat file. Files with the .mat extension are the binary data container format that are used by the MATLAB program. The extension was developed by Mathworks and MAT files are categorized as data files that include variables, functions, arrays and other information and we need the access in python since we are working on python. To access the segmentation data from mat file we have used mat73 library in python. Mat73 library loads MATLAB 7.3 HDF5 files into a Python dictionary for futher processing.

# 

# 3.2 Radiomics Limitations

These days, difference in radiomics workflow among different researchers lead to different results which cause compromising the reproducibility of the technique. Additionally, this makes it more tough to establish a validated and useful radiomic signature.

## **3.2.1 Imaging Acquisition**

The type of machine used throughtout the image acquisition, the different slice thickness, reconstruction matrices, configurations, fields of view, the time point of the administration and type of contrast and also the time point of the investigation itself during the course of illness might result in relevant differences in the image. Of these parameters have an crucial correlation with the VOI sizes and therefore the ROI definition during segmentation affecting radiomic markers [38].

## **3.2.2 Image Segmentation**

Image segmentation represents one in all the foremost necessary steps throughout radiomic progress. It is performed manually, semiautomatically or automatically. Manual segmentation represents a time-consuming procedure. This also needs to be carried out by a specialist, and a high interobserver variability can be expected. Automatic segmentation represents a more reproducible and quicker way to do segmentation and is incredibly helpful for giant imaging datasets. Semiautomatic segmentation requires an interaction between the observer and software. Most often, the observer has to outline the VOI and will refine the semiautomatic segmentation results manually. Since our dataset is small we have manually defined the segmentation in RTSS.mat file. There exist some automatic and semiautomatic segmentation methods however they are not suitable for every VOI and thus need to be upgraded for certain problems and used with standardized adjustments when they are applied [38].

## **3.2.3 Feature Extraction**

Imaging features are extracted from the VOI areas outlined by the segmentation. each feature represents one half from the complete radiomic concept. The tumor intensity is quantified by an intensity histogram that displays the three-dimensional fractional volume data for the range of voxel values. Shape information contain values just like the total volume, surface area, compactness, and actual form of the lesion. Texture based features are put together by mathematical algorithms that deliver second order statistics or co-occurrent matrix features. Many of values could also be calculated containing information related to additional qualities like densities, homogeneity, grey shades, clusters, correlations between those in different settings, and much more.

A common challenge in radiomics is to outline a non-redundant set of imaging biomarkers from the huge number of extracted features, and to boost the radiomic performance, redundancies should be excluded [38]. After feature extraction, the use of some statistical methods or supporting our workflow with machine learning allows us to select the most precise features reducing the spectrum to those validated.

## **3.2.4 Image Processing**

To make a successful image processing procedure, an appropriate bioinformatics approach must be chosen that is able to cope with big data, to obtain a reproducible and solid biomarker signature. However, this make it necessary to perform careful data management to improve the study results and diminish the influence of outliers, blurring values, interobserver variability, and diffuse readings [39]. A number of radiomic features can be extracted from each tumor despite having as many patients as possible however they should ideally be integrated into the study. This is of course often limited even in multicenter studies. Generally, those relevant markers are determined in retrospective cohort studies. When a radiomic signature has been identified, it should be verified in an independent cohort in a prospective study [40].

## Chapter 4

# Experimental results

Our aim in this research paper is to extract the various features from the head and neck cancer patient’s dataset and combine them in a single file for futher processing. As we get the one csv file containing all features we need to use feature selection and then feed this data into machine learning algorithm. Both processes are explained briefly in next chapter Future direction.

During implementation of the code I faced lots of Challenges ,Here I will mention few of them here,

1. Conversion of DICOM files into nparray file
2. Extraction of contour segmentation from RTSS.mat file, since there are different methods for loading mat file in python ,however for matlab version 7.3 old python libraries are not working.
3. Getting only images which has contours in RTSS.mat file
4. Masking the binary mask over the image
5. Different libraries and methods are available for extracting the features
6. Putting all features in a csv file
7. Concatenate all csv files as one

I came up with 3 different implantation or extraction strategies the first was to extract the features from raw dicom image just for visualization and can compare the result with other methods later. Below given code extract the features from dicom images directly. This code extract the feature with the id’s of patients to see which feature is extracted from which patient and its regarding image.

#features for all dicom images for experiment purpose

dataframe = get\_df\_from\_img\_array(n\_array,ids, getId = True)

print("Extracted features for all images : ")

df = pd.DataFrame(dataframe)

df.to\_csv('1.csv')

'''

Below given code extract the feature from images without id’s, these the raw features which will be feed into machine learning algorithm without id’s. Since there is no need to train a model with patient’s id.

'''

#from dcm images features without id for experiment purpose

dfa = get\_df\_from\_path(data\_path , [])

print("5 df from top are: ")

df = pd.DataFrame(dfa)

df.to\_csv('output\_csv\\'+str(folders[f])+'.csv')

'''

Please refer to fig. 1 to see the dicom image and to see the feature Please refer to 1.csv after running the code. Since features consist lots of data which could not be visualize all of them here.

Second was to convert the dicom and mask from RTTS.mat to jpg and overlapping these two jpg together. The results obtained from this method where very poor and the extracted information didn’t have much use since it was simple from jpg . Below mentioned code give the overlapped image and extract the feature from it.

masks\_files = glob(masks\_path+'\*.jpg')

jpg\_files = glob(jpg\_path+'\*.jpg')

for i in range(0,len(masked\_files)):

img=cv2.imread(masks\_files[i])

im2=cv2.imread(jpg\_files[i])

dst = cv2.addWeighted(img, 0.5, im2, 0.5, 0)

plt.imsave(overlayed\_path+str(i)+'overlayed.jpg',dst)

Y\_data = []

overlayed\_files=glob(overlayed\_path+'\*.jpg')

for myFile in overlayed\_files:

overlayed\_image = cv2.imread (myFile)

gray\_image = cv2.cvtColor(overlayed\_image, cv2.COLOR\_BGR2GRAY)

Y\_data.append (gray\_image)

dataframe = get\_df\_from\_img\_array(Y\_data,ids, getId = True)

print("Extracted features for overlayed images : ")

df = pd.DataFrame(dataframe)

df.to\_csv('output\_csv\\'+str(folders[f])+'.csv')

For detailed code please refer to appendix B. Fig 3 is the overlayed image of patient image and mask.

I didn’t settle for this because I felt extraction from this wasn’t specific to the ROI the results from this are reliable if we just wanted to extract general feature but in this project we need a specific region I couldn’t take a chance.

Finally I decided to use this method I called masked array because in order to handle the data correctly in python at some points in the code we turn it into dictionaries or arrays, this involved multiplying the binary information in the mat file and the dicom file ,Since we have binary mask which consist 0’s and 1’s and we have converted the dicom data to array to so we can mask both of them to get masked image, Size of the both arrays are same so we don’t need to resize them. And masking is just multiplying the both arrays so we get the ROI and can extract the features of it. For extracting feature from the Masked data, I convert it into grayscale image and extract the feature from it. Below given code extract the features from masked image.

X\_data = []

masked\_files = glob(masked\_path+'\*.jpg')

for myFile in masked\_files:

masked\_image = cv2.imread (myFile)

gray\_image = cv2.cvtColor(masked\_image, cv2.COLOR\_BGR2GRAY)

X\_data.append (gray\_image)

#print('X\_data shape:', np.array(X\_data).shape)

dataframe = get\_df\_from\_img\_array(X\_data,ids, getId = True)

df = pd.DataFrame(dataframe)

print("Extracted features for masked images for patient's ID : "+folders[f])

df.to\_csv('output\_csv\\'+str(folders[f])+'.csv')

For detailed code please refer to appendix B. Fig 4 is the masked image. To visualize output please refer to fig. 5 and 7.

the only drawback to this was the small possibility of loss of data between changing datatypes but the results acquired from this were the most appropriate .

The first script of my code required the user to manually create 4 folders for each patient this was done for experiment and visualization and my own understanding the script . I created four empty folders in each patient’s folder named ‘image’ which will contain jpg CT images, ‘masks’ which will contain jpg masks , ‘masked’ which will contain masked jpg images and ‘overlayed’ which will contain combined jpg images of CT images and masks images for visualization. I am simply extracting all the data and saving them in image format in their respective folders.

The second script I wrote implemented 3rd method of feature extraction and required little input from user and it does not require to create 4 different folders manually since dataset was small and I wanted to differenciate the result visually I had written the first script , If dataset is large then we can not manually create the folder for each patient so to resolve this problem I came with second script. In this script I created a single common folder in which I save the masked images for first patient and extract the features from them and then empty the folder for second patient and it happens resursively until we are able to extract features for all patients. As we are able to extract the features for all patients we have multiple csv files saved in another folder in this script I managed to combine all of them into one within that folder. So overall I have to create only two folders. Below code combine all the csv file by taking the path where all csv files are stored and combined them as final.csv.

final = 'C:\\Users\\peace\\Desktop\\12\\output\_csv\\\*.csv'

df = pd.concat((pd.read\_csv(f) for f in glob(final, recursive=True)), ignore\_index=True)

df.to\_csv('final.csv')

Please refer to appendix B for full code. Refer to fig. 5 and 6 to see the outcome of the code.

Due to time I couldn’t carry on the radiomic workflow all the way to a working a svm model so I decided it was best put labels on the results but I managed to code for removing them and I left comments in the script. However I found this topic very interesting and I will keep researching on it

# Chapter 5

# Conclusions And Future Direction

## **5.1 Conclusions**

This paper concludes that there is great potential for radiomics and texture analysis techniques for improving upon many aspects of the tumor assessment, risk stratification, and outcome evaluation aspects in head and neck cancer therapy. An attempt towards standardization of radiomics algorithms and specific acquisition parameters is critically required for the oncologic community to outline the role of radiomics and texture analysis techniques in a manner that the clinical practice takes in. New studies regarding specific head and neck sublocalization are required.

## **5.2 Future Direction**

In the future, the work of this thesis research can be extended to increase the detection and segmentation accuracy using automatic segmentation. In this thesis I have extracted the segmented image of patients and on the basis of that extracted the different features. In the future we can also use feature selection methods so that we can group the best features for better output and then feed them into the machine learning model for head and neck cancer detection.

## **5.2.1 Feature Selection**

Feature selection is meant here to refer to the problem of dimensionality reduction of data, which initially contain a high number of features. One hopes to choose optimal subsets of the original features that still contain the information essential for the classification task, while reducing the computational burden imposed by using many features.

There are many reasons for using feature selection technique to reduce the number of features.

* Satisfying the general goal of maximizing the accuracy of the classifier while minimizing the associated measurement costs.
* Improving accuracy by reducing irrelevant and possibly redundant features.
* Reducing the complexity and the associated computational cost.
* Reduce the amount of data needed for the training.
* Improving the chances that a solution will be both understandable and practical.

There are different feature selection algorithms which may help to find the set of features which will provide the best solutions. algorithms such as Genetic Algorithm, Particle Swarm Algorithm, Ant Colony Optimization and Artificial Bee Colony Optimization algorithm are proposed for feature selection.

## **5.2.2 Machine learning algorithms**

Machine learning can be broadly defined as the collection of computational methods using data to improve performance or make accurate prediction. These programmable methods can learn from the data, and hence automate and improve the prediction process. Recently Parmar et al. assessed a large panel of machine-learning methods for overall survival prediction of head and neck cancer patients. They investigated 12 machine-learning classifiers belonging to the 12 classifier families: Bagging (BAG), Bayesian (BY), boosting (BST), decision trees (DT), discriminant analysis (DA), generalized linear models (GLM), multiple adaptive regression splines (MARS), nearest neighbors (NN), neural networks (Nnet), partial least square and principle component regression (PLSR), random forests (RF), and support vector machines (SVM). In this study the authors practically demonstrated and additionally showed high prognostic performance and stability of machine learning methods applied to Radiomics features [41].

Thesedays, radiomic features can predict some tumor characteristics related to survival in some head and neck cancer [42]. additionally, some info from the study of Aerts et al. [43] in non-small cell lung cancer suggest that when the TNM classification is combined with radiomics signature, the performance is considerably better than TNM alone, suggesting complementary data for prognos

## **References**

[1] Global Burden of Disease Cancer Collaboration, Fitzmaurice C, Allen C, et al. Global, Regional, and National Cancer Incidence, Mortality, Years of Life Lost, Years Lived With Disability, and Disability-Adjusted Life-years for 32 Cancer Groups, 1990 to 2015: A Systematic Analysis for the Global Burden of Disease Study [published correction appears in JAMA Oncol. 2017 Mar 1;3(3):418]. JAMA Oncol. 2017;3(4):524‐548. doi:10.1001/jamaoncol.2016.5688

[2] Lu G, Little JV, Wang X, et al. Detection of Head and Neck Cancer in Surgical Specimens Using Quantitative Hyperspectral Imaging. Clin Cancer Res. 2017;23(18):5426‐5436. doi:10.1158/1078-0432.CCR-17-0906

[3] Pan R, Zhu M, Yu C, et al. Cancer incidence and mortality: A cohort study in China, 2008-2013. Int J Cancer. 2017;141(7):1315‐1323. doi:10.1002/ijc.30825

[4] Head and neck cancer Dr Athanassios Argiris, Michalis V Karamouzis, David Raben, Robert L Ferris, Published:May 17, 2008DOI:https://doi.org/10.1016/S0140-6736(08)60728-X

[5] Epidemiology of Human Papillomavirus–Positive Head and Neck Squamous Cell Carcinoma Maura L. Gillison, Anil K. Chaturvedi, William F. Anderson, and Carole Fakhry

Journal of Clinical Oncology 2015 33:29, 3235-3242

[6] Helliwell, T., & Giles, T. (2016). Pathological aspects of the assessment of head and neck cancers: United Kingdom National Multidisciplinary Guidelines. The Journal of Laryngology & Otology, 130(S2), S59-S65. doi:10.1017/S0022215116000451

[7] National Comprehensive Cancer Network. NCCN Clinical Practice Guidelines in Oncology (NCCN Guidelines®): Head and Neck Cancers

[8] Chinese Society of Clinical Oncology (CSCO) diagnosis and treatment guidelines for head and neck cancer 2018 (English version). Chin J Cancer Res. 2019;31(1):84‐98. doi:10.21147/j.issn.1000-9604.2019.01.05

[9] Clinical Practice in PET/CT for the Management of Head and Neck Squamous Cell Cancer Reema Goel, William Moore, Baran Sumer, Saad Khan, David Sher, and Rathan M. Subramaniam American Journal of Roentgenology 2017 209:2, 289-303

[10] Gao S, Li S, Yang X, Tang Q. 18FDG PET-CT for distant metastases in patients with recurrent head and neck cancer after definitive treatment. A meta-analysis. Oral Oncol. 2014;50(3):163‐167. doi:10.1016/j.oraloncology.2013.12.002

[11] Positron Emission Tomography With [18F]Fluorodeoxyglucose Improves Staging and Patient Management in Patients With Head and Neck Squamous Cell Carcinoma: A Multicenter Prospective Study

Max Lonneux, Marc Hamoir, Hervé Reychler, Philippe Maingon, Christian Duvillard, Gilles Calais, Boumédiène Bridji, Laurence Digue, Michel Toubeau, and Vincent Grégoire

Journal of Clinical Oncology 2010 28:7, 1190-1195

[12] https://www.nccn.org/professionals/physician\_gls/default.aspx

[13] Mullins ME. Modern emergent stroke imaging: pearls, protocols, and pitfalls. Radiol Clin North Am. 2006;44(1):41‐viii. doi:10.1016/j.rcl.2005.08.002

[14] Histogram-Based Texture Characterization and Classification of Brain Tissues in Non-Contrast CT Images of Stroke Patients By Kenneth K. Agwu and Christopher C. Ohagwu, Submitted: March 17th 2016Reviewed: August 24th 2016Published: December 14th 2016;DOI: 10.5772/65349

[15] Das, Swagatam & Abraham, Ajith & Konar, Amit. (2006). Spatial Information Based Image Segmentation Using a Modified Particle Swarm Optimization Algorithm. 10.1109/ISDA.2006.253877.

[16] Behera, Himansu Sekhar. “Segmentation and Classification using Heuristic HRSPSO.” (2011).

[17] https://www.intechopen.com/books/pattern-recognition-analysis-and-applications/histogram-based-texture-characterization-and-classification-of-brain-tissues-in-non-contrast-ct-imag

[18] https://pyradiomics.readthedocs.io/en/latest/\_modules/radiomics/firstorder.html

[19] https://pyradiomics.readthedocs.io/en/latest/features.html#module-radiomics.glcm

[20] https://www.spcforexcel.com/knowledge/basic-statistics/are-skewness-and-kurtosis-useful-statistics

[21] Castellano G, Bonilha L, Li LM, Cendes F. Texture analysis of medical images. Clin Radiol 2004; 59 (12): 1061–1069.

[22] https://medium.com/@thevatsalsaglani/feature-extraction-from-medical-images-and-an-introduction-to-xtract-features-9a225243e94b

[23] https://buildmedia.readthedocs.org/media/pdf/pyradiomics/v1.0.1/pyradiomics.pdf

[24] http://www.ajnr.org/content/suppl/2011/07/12/ajnr.A2232.DC1/2232\_table.pdf

[25] Role of the texture features of images in the diagnosis of solitary pulmonary nodules in different sizes. Qian Zhao, Chang-Zheng Shi, Liang-Ping Luo, Department of Statistics, School of Public Health, Guangzhou Medical University, Guangzhou 510182, China; Medical Imaging Center, the First Affiliated hospital of Jinan University, Guangzhou 510630, China Correspondence to: Liang-Ping Luo, MD. Medical Imaging Center, First Affiliated Hospital, Jinan University, No.613, Huangpu west Rd, Guangzhou 510630, China. Email: tluolp@jnu.edu.cn.

[26] https://www.learnopencv.com/shape-matching-using-hu-moments-c-python/

[27] https://in.mathworks.com/help/images/pixel-values-and-image-statistics.html

[28] https://github.com/ninn55/python-regionprops

[29] https://in.mathworks.com/help/images/ref/regionprops.html

[30] https://towardsdatascience.com/types-of-convolution-kernels-simplified-f040cb307c37

[31] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6322843/figure/F1/

[32] http://tcr.amegroups.com/article/view/8805/html

[33] Radiomics and Texture Analysis in Laryngeal Cancer. Looking for New Frontiers in Precision Medicine through Imaging Analysis by Carlos Miguel Chiesa-Estomba ,Oier Echanizl Ekhiñe Larruscain ,Jose Angel Gonzalez-Garcia,Jon Alexander Sistiaga-Suarez, andManuel Graña

[34] hang, L.; Fried, D.V.; Fave, X.J.; Hunter, L.A.; Yang, J.; Court, L.E. IBEX: An open infrastructure software platform to facilitate collaborative work in radiomics. Med. Phys. 2015, 42, 1341–1353.

[35] Brown, A.M.; Nagala, S.; McLean, M.A.; Lu, Y.; Scoffings, D.; Apte, A.; Gonen, M.; Stambuk, H.E.; Shaha, A.R.; Tuttle, R.M.; et al. Multi-institutional validation of a novel textural analysis tool for preoperative stratification of suspected thyroid tumors on diffusion-weighted MRI. Magn. Reson. Med. 2016, 75, 1708–1716.

[36] Fang, Yu-Hua & Lin, Chien-Yu & Shih, Meng-Jung & Wang, Hung-Ming & Ho, Tsung-Ying & Liao, Chun-Ta & Yen, Tzu-Chen. (2014). Development and Evaluation of an Open-Source Software Package “CGITA” for Quantifying Tumor Heterogeneity with Molecular Images. BioMed research international. 2014. 248505. 10.1155/2014/248505.

[37] Zhang, H.; Graham, C.M.; Elci, O.; Griswold, M.E.; Zhang, X.; Khan, M.A.; Pitman, K.; Caudell, J.J.; Hamilton, R.D.; Ganeshan, B.; et al. Locally advanced squamous cell carcinoma of the head and neck: CT texture and histogram analysis allow independent prediction of overall survival in patients treated with induction chemotherapy. Radiology 2013, 269, 801–809.

[38] Kumar, V.; Gu, Y.; Basu, S.; Berglund, A.; Eschrich, S.A.; Schabath, M.B.; Forster, K.; Aerts, H.J.; Dekker, A.; Fenstermacher, D.; et al. Radiomics: The process and the challenges. Magn. Reson. Imaging 2012, 30, 1234–1248.

[39] Aerts, H.J.; Velazquez, E.R.; Leijenaar, R.T.; Parmar, C.; Grossmann, P.; Carvalho, S.; Bussink, J.; Monshouwer, R.; Haibe-Kains, B.; Rietveld, D.; et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat. Commun. 2014, 5, 4006, Erratum in: Nat. Commun. 2014, 5, 4644.

[40] Gillies, R.J.; Kinahan, P.E.; Hricak, H. Radiomics: Images Are More than Pictures, They Are Data. Radiology 2016, 278, 563–577.

[41] armar, C.; Grossmann, P.; Rietveld, D.; Rietbergen, M.M.; Lambin, P.; Aerts, H.J. Radiomic Machine-Learning Classifiers for Prognostic Biomarkers of Head and Neck Cancer. Front. Oncol. 2015, 5, 272.

[42] Giraud, P.; Giraud, P.; Gasnier, A.; El Ayachy, R.; Kreps, S.; Foy, J.P.; Durdux, C.; Huguet, F.; Burgun, A.; Bibault, J.E. Radiomics and Machine Learning for Radiotherapy in Head and Neck Cancers. Front. Oncol. 2019, 9, 174.

[43] Alic, L.; Niessen, W.J.; Veenland, J.F. Quantification of heterogeneity as a biomarker in tumor imaging: A systematic review. PLoS ONE 2014, 9, e110300.

[44] https://www.cancer.net/cancer-types/head-and-neck-cancer/view-all

[45] https://www.cancer.net/cancer-types/head-and-neck-cancer/statistics

[46] https://www.cancer.net/cancer-types/head-and-neck-cancer/risk-factors-and-prevention

[47] https://www.cancer.net/cancer-types/head-and-neck-cancer/symptoms-and-signs

## **ACKNOWLEDGEMENTS**

I am heartily thankful to my mentor Lecturer Hongxia Hao for her appreciative sustain guidance and support to me during the time period of my literature review. Her constant motivation and expert opinions, at many intervals, helped me to shape and accomplish this research work. Without her valuable and timely suggestions and observations, this research work would not have been probably completed in the present form.

I am also thankful to this year’s graduation committee who have made sure to provide me with all the necessary support to make sure that I managed to get my work on time during a year were the world was affected by an epidemic, these man and women have my deepest gratitude for working extra hours to be available when I needed their services. On the part of being available and offering help I would also like to thank the Xidian University SIE administration who worked tireless to keep us informed and in touch with our Tutors.

I am thankful to my family who were supportive of me during my research work and offering words of motivation.

# Appendix A

Cancer begins once healthy cells change and grow in size out of control, forming a unwanted mass called a tumor. A tumor could be cancerous or benign. A cancerous tumor is malignant, meaning it is dangerous and can grow and spread to other parts of the body. A benign tumor means the tumor could grow however will not spread[44].

**A1 About head and neck cancer**

Head and neck cancer consist a number of different malignant tumors that grow in or around the throat, larynx, nose, sinuses, and mouth. Most head and neck cancers are squamous cell carcinomas. This type of cancer begins within the flat squamous cells that create the thin layer of tissue on the surface of the structures in the head and neck. Directly to the lower place of this lining, which is called the epithelium, some areas of the head and neck have a layer of moist tissue, referred as the mucosa. If a cancer is simply found within the squamous layer of cells, it is referred to as carcinoma in situ. If the cancer is growing beyond this cell layer and reached down into the deeper tissue, then it is called invasive squamous cell carcinoma. If doctors cannot determine wherever the cancer began, it's referred to as a cancer of unknown primary.

* **Laryngeal and hypopharyngeal cancer.** The larynx is commonly referred as the voice box and The tube shaped organ in the neck is necessary for breathing, talking, and swallowing(fig. 6). It is located at the top of the windpipe, or trachea. The hypopharynx is also called the gullet. It is the lower part of the throat that surrounds the larynx.
* **Nasal cavity and paranasal sinus cancer.** The nasal cavity is the part just behind the nose where air passes through throat. The paranasal sinuses are the air-filled areas that surround the nasal cavity.
* **Nasopharyngeal cancer.** The nasopharynx is the air passing way at the upper part of the throat behind the nose.
* **Oral and oropharyngeal cancer.** The oral cavity includes the mouth and tongue. The oropharynx consist the middle of the throat, from the tonsils to the end of the voice box.
* **Salivary gland cancer.** The salivary gland produces saliva and Saliva is the liquid that is released into the mouth to keep it moist and this fluid contains enzymes that helps in breaking down the food.

**A2 Head and Neck Cancer: Statistics**

survival rates rely upon several factors, Head and neck cancer accounts for around 4% of all cancers in the United States. This year, an estimated 65,630 people which include 48,200 men and 17,430 women will develop head and neck cancer. It is estimated that 14,500 deaths which include 10,760 men and 3,740 women from head and neck cancer will occur this year. The 5 year survival rate is the percent of people live at least 5 years after the cancer is discovered. Here Percent means how many out of 100. The 5 year survival rate for individuals with head and neck cancer varies and depends on different factors. It is important to remember that statistics on the survival rates for people with head and neck cancer are an estimate. The estimate comes from annual data based on the number of individuals with head and neck cancer in the United States. Also, specialists calculate the survival statistics every 5 years. So the estimate 5 year survival rate may not show the results of better diagnosis or treatment available for less than 5 years. Statistics adapted from the American Cancer Society’s publication, Cancer Facts & Figures 2020, and the National Cancer Institute (January 2020) [45].

**A3 Head and Neck Cancer: Medical Illustrations**

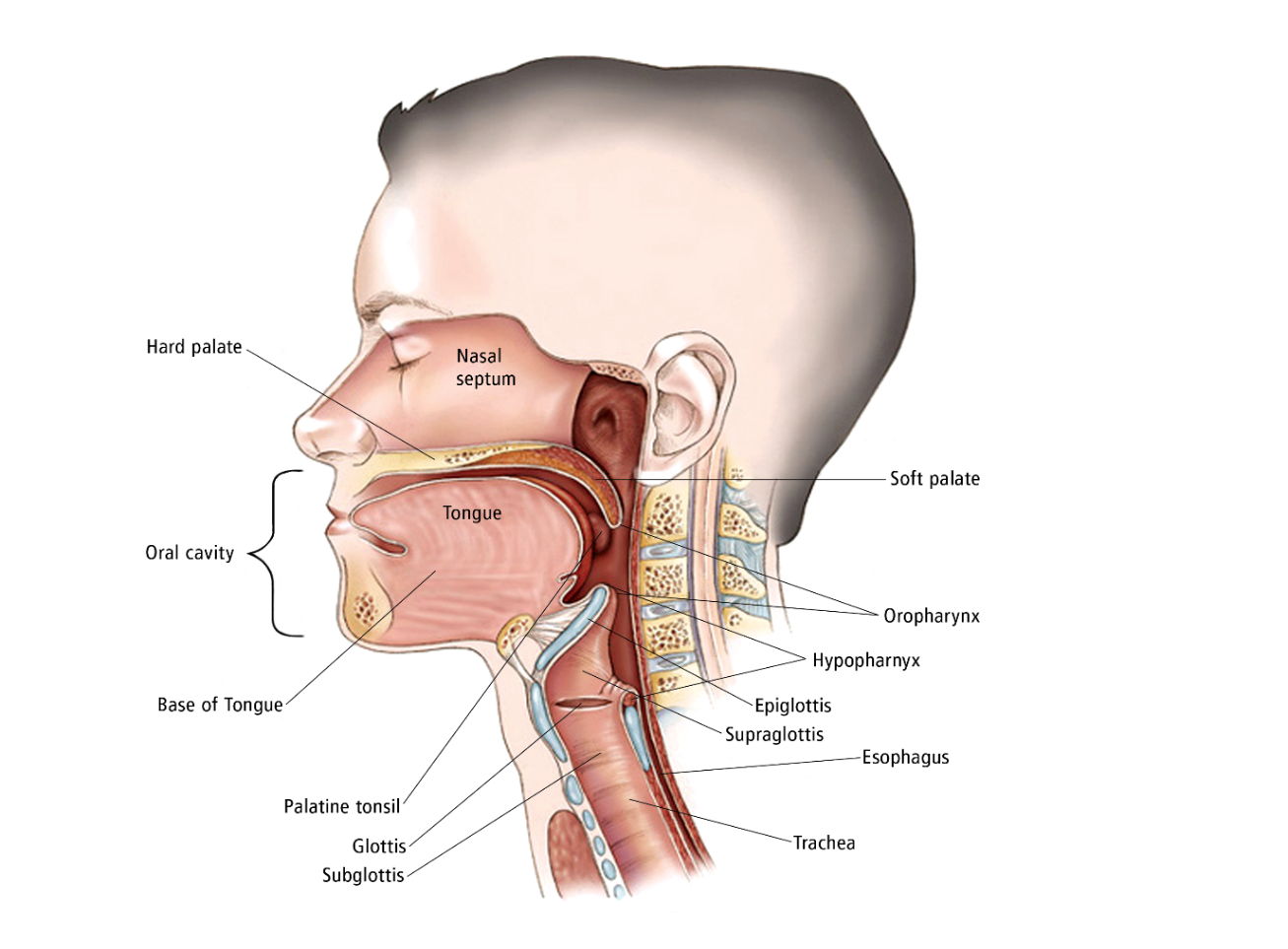


Fig 9. Drawing of the main body parts affected by head and neck cancer.

**A4 Risk Factors and Prevention**

**A4.1 Risk Factors**

A risk factor is something that will increases a person’s likelihood of developing cancer. Though risk factors typically influence the development of cancer, mostly do not directly cause cancer. Some people with many risk factors never develop the disease, whereas others with no known risk factors can develop the cancer. Knowing your risk factors and concerning them together with your doctor could assist you make more informed lifestyle and health care choices.

There are two substances that cause drastically increase the risk of developing a head and neck cancer:

* **Tobacco.** Tobacco use includes smoking cigarettes, cigars, or pipes; chewing tobacco; and using snuff. Tobacco is the single largest risk factor for head and neck cancer. 85% of head and neck cancers are related to tobacco use, and the amount of tobacco use may affect prognosis, which is the chance of recovery. Additionally, second hand smoke might increase a individual’s risk of developing head and neck cancer.
* **Alcohol.** Frequent and heavy alcohol consumption increase the chances of developing cancer in the mouth, pharynx, larynx, and esophagus.  
  Using alcohol and tobacco together will increases the risk of developing cancer.

Other factors that can raise a individual’s risk of developing head and neck cancer include:

* **Prolonged sun exposure.** This is related to cancer in the lip area, as well as skin cancer of the head and neck.
* **Human papillomavirus (HPV).** Research demonstrates that infection with HPV is a risk factor causing head and neck cancer. HPV can transmit through one person to another person during sexual activities. There are many types of HPV, known as strains. Research connects some HPV strains more strongly with certain types of cancers. There are vaccines available to protect you from the HPV strains that cause head and neck cancer.
* **Epstein-Barr virus (EBV).** Direct exposure to EBV, which is more commonly referred as the virus that causes mononucleosis or "mono," plays a important role in the development of nasopharyngeal cancer.
* **Gender.** Men are more prone to the head and neck cancer than women. Men are 2 to 3 times more likely than women to develop head and neck cancer. But, the rate of head and neck cancer in women has been increasing for several decades.
* **Age**. Peole of age 40 or above are at the higher risk of having head and neck cancer.
* **Poor oral and dental hygiene.** Poor oral care may increase the risk of head and neck cancer.
* **Environmental or occupational inhalants.** Inhaling asbestos, wood dust, paint fumes, and certain chemicals may increase a individual’s chances of head and neck cancer.
* **Marijuana use**. Research suggests that people using marijuana may be at higher risk for head and neck cancer.
* **Poor nutrition.** A diet with low vitamins A and B can raise a person’s chance of head and neck cancer.
* **Gastroesophageal reflux disease (GERD) and laryngopharyngeal reflux disease (LPRD).** Reflux of stomach acid into the upper airway and throat may be related with the development of head and neck cancer.
* **Weakened immune system**. A weak immune system can raise a individual’s risk of head and neck cancer.
* **Exposure to radiation.** It is associated with salivary gland cancer.
* **Previous history of head and neck cancer.** People having head and neck cancer have a higher risk of developing another head and neck cancer in the future.

**A4.2 Prevention**

There Different factors which cause different types of cancer. Researchers still scrutinize what factors cause this kind of cancer, including with ways in which to stop it. Although there is no proven way to completely prevent this disease however you can lower your risk [46].

Stopping the use of all tobacco product is that the most vital issue an individual will do to cut back their risk, even for those who are smoking for several years. Other steps that can reduce the risk of head and neck cancer include:

* Avoiding alcohol
* avoiding marijuana use
* Using sunscreen regularly, including lip balm with an adequate sun protection factor (SPF)
* Reducing your risk of HPV infection by taking the HPV vaccine or by limiting your number of sexual partners, having many partners increases the risk of HPV infection. Also using a condom during sexual activity cannot fully protect you from HPV.
* Maintaining proper oral care.

**A5 Head and Neck Cancer: Symptoms and Signs**

People with head and neck cancer typically expertise the subsequent symptoms or signs. Sometimes, folks with head and neck cancer don't have any of those changes. Or, the reason behind a symptom could also be a unique medical condition that's not cancer[47].

* Swelling or a sore that does not heal; this is the most common symptom
* Red or white patch in the mouth
* Lump, bump, or mass in the head or neck area, with or without pain
* Persistent sore throat
* Foul mouth odor not explained by hygiene
* Hoarseness or change in voice
* Nasal obstruction or persistent nasal congestion
* Frequent nose bleeds and/or unusual nasal discharge
* Difficulty breathing
* Double vision
* Numbness or weakness of a body part in the head and neck region
* Pain or difficulty chewing, swallowing, or moving the jaw or tongue
* Jaw pain
* Blood in the saliva or phlegm, which is mucus discharged into the mouth from respiratory passages
* Loosening of teeth
* Dentures that no longer fit
* Unexplained weight loss
* Fatigue
* Ear pain or infection

# 

# 

# Appendix B

**Code:**

import mat73

import cv2

import pandas as pd

import numpy as np

import pydicom as dicom

import dicom

import os

import matplotlib.pyplot as plt

from glob import glob

from xtract\_features.helpers import \*

from xtract\_features.glcms import \*

from xtract\_features.extract import get\_df\_from\_img\_array

from xtract\_features.extract import feature\_dict\_from\_imgarray, get\_df\_from\_path

from xtract\_features.extract import s\_entropy, entropy\_simple

from xtract\_features.twodconv import conv2d

import scipy.ndimage as ndi

from scipy.stats import skew

from scipy.stats import kurtosis

import warnings

warnings.filterwarnings('ignore')

#Testing start

data\_path = "C:\\Users\\peace\\Desktop\\12\\HGJ\\HN-HGJ-001\\CT\\image\\"

rtss\_path='C:\\Users\\peace\\Desktop\\12\\HGJ\\HN-HGJ-001\\CT\\RTS\\'

g = glob(data\_path + '/\*.dcm')

data\_dict = mat73.loadmat(rtss\_path+'RTSS.mat')

data = np.array(data\_dict)

a=data\_dict['contours']['Segmentation'].astype(np.uint8)

lst\_numpy\_arrs, ids = extract\_img\_array(data\_path, getID=True)

print(a.shape)

print(ids[:10])

#Convert all dicom images to nparray and putting all arrays to npy file

def load\_scan(path):

slices = [dicom.read\_file(path + '/' + s) for s in os.listdir(path)]

slices.sort(key = lambda x: int(x.InstanceNumber))

try:

slice\_thickness = np.abs(slices[0].ImagePositionPatient[2] - slices[1].ImagePositionPatient[2])

except:

slice\_thickness = np.abs(slices[0].SliceLocation - slices[1].SliceLocation)

for s in slices:

s.SliceThickness = slice\_thickness

return slices

def get\_pixels\_hu(scans):

image = np.stack([s.pixel\_array for s in scans])

# Convert to int16 (from sometimes int16),

# should be possible as values should always be low enough (<32k)

image = image.astype(np.int16)

# Set outside-of-scan pixels to 1

# The intercept is usually -1024, so air is approximately 0

image[image == -2000] = 0

# Convert to Hounsfield units (HU)

intercept = scans[0].RescaleIntercept

slope = scans[0].RescaleSlope

if slope != 1:

image = slope \* image.astype(np.float64)

image = image.astype(np.int16)

image += np.int16(intercept)

return np.array(image, dtype=np.int16)

id=0

patient = load\_scan(data\_path)

imgs = get\_pixels\_hu(patient)

np.save("fullimages\_%d.npy" % (id), imgs)

file\_used="fullimages\_%d.npy" % id

n\_array = np.load(file\_used).astype(np.float64)

#show every 3rd image from 10th image from the npy file

def sample\_stack(stack, start\_with=10,rows=6,cols=6, show\_every=3):

fig,ax = plt.subplots(rows,cols,figsize=[12,12])

for i in range(rows\*cols):

ind = start\_with + i\*show\_every

ax[int(i/rows),int(i % rows)].set\_title('slice %d' % ind)

ax[int(i/rows),int(i % rows)].imshow(stack[ind],cmap='gray')

ax[int(i/rows),int(i % rows)].axis('off')

plt.show()

sample\_stack(n\_array)

plt.hist(n\_array.flatten(), bins=50, color='c')

plt.xlabel("Hounsfield Units (HU)")

plt.xlim(-2000, 2000)

plt.ylabel("Frequency")

plt.show()

#one Example of masked image and its features that we are extracting

c=a[72]\*n\_array[72]

plt.imshow(a[72])

plt.show()

print("Mean intensity")

print(cv2.mean(c))

print("varience")

print(np.var(c))

print("skewness")

print(skew(c, axis=0, bias=True))

print("kurtosis")

print(kurtosis(c, axis=0, bias=True))

print("Shannon's entropy for image ")

print(s\_entropy(c))

print("Simple entropy for image ")

print(entropy\_simple(c))

feats = glcm(c)

# correlation

corr = feats.correlation()

print("correlation ")

print(corr)

# homogeneity

homogeneity = feats.homogeneity()

print("homogeneity ")

print(homogeneity)

# contrast

cont = feats.contrast()

print("contrast")

print(cont)

# energy

energy = feats.energy()

print("energy ")

print(energy)

'''

# all glcm features at once

allf = feats.glcm\_all()

print("array masked image ")

print(allf)

'''

"""Image moments are used to describe objects after segmentation and play an

essential role in object recognition and shape analysis. Images moments may be

employed for pattern recognition in images. Simple image properties derived via

raw moments is area or sum of grey levels. \_moments is a list of 24 variant

moments and \_hu\_moments is list of the 7 hu moments which are invariant."""

from xtract\_features.moments import \*

hu\_moments = moments(c).get\_HuMoments()

print("printing hu moments")

print(hu\_moments)

moments = moments(c).get\_moments()

print("printing moments")

print(moments)

from xtract\_features.region\_props import \*

rp = region\_props(c)

# maximum area region

max\_area = rp.max\_area()

print("printing max\_area")

print(max\_area)

# plot black and white

rp.plot\_show\_bw()

print("plot black and white")

plt.show()

# mean of areas of all the regions

rp.mean\_area()

print("mean of areas of all the regions")

print(rp.mean\_area())

# eccentricity of the highest area region

rp.eccentricity()

print("eccentricity of the highest area region")

print(rp.eccentricity())

rp.euler\_number()

print("rp.euler\_number()")

print(rp.euler\_number())

rp.solidity()

print("rp.solidity()")

print(rp.solidity())

rp.perimeter()

print("rp.perimeter()")

print(rp.perimeter())

# standard deviation of all the areas of the regions of the given image

rp.std\_area()

print("rp.std\_area()")

print(rp.std\_area())

# otsu’s Threshold

rp.thresh\_img()

print("erp.thresh\_image()")

print(rp.thresh\_img())

rp.bb()

print("rp.bb()")

print(rp.bb())

rp.bb\_area()

print("rp.bb\_area()")

print(rp.bb\_area())

rp.centroid\_r()

print("rp.centroid\_r()")

print(rp.centroid\_r())

rp.convex\_area\_r()

print("rp.convex\_area\_r()")

print(rp.convex\_area\_r())

rp.coordinates\_r()

print("rp.coordinates\_r()")

print(rp.coordinates\_r())

rp.eq\_diameter()

print("rp.eq\_diameter()")

print(rp.eq\_diameter())

rp.extent\_r()

print("rp.extent\_r()")

print(rp.extent\_r())

rp.filled\_area\_r()

print("rp.filled\_area\_r()")

print(rp.filled\_area\_r())

rp.inertia\_tensor\_r()

print("rp.inertia\_tensor\_area()")

print(rp.inertia\_tensor\_r())

rp.label\_r()

print("erp.label\_r()")

print(rp.label\_r())

rp.inertia\_tensor\_eigvals\_r()

print("rp.inertia\_tensor\_eigvals\_r()")

print(rp.inertia\_tensor\_eigvals\_r())

rp.local\_centroid\_r()

print("rp.local\_centroid\_r()")

print(rp.local\_centroid\_r())

rp.maj\_ax\_len()

print("rp.maj\_ax\_len()")

print(rp.maj\_ax\_len())

rp.min\_ax\_len()

print("rp.min\_ax\_len()")

print(rp.min\_ax\_len())

rp.orient()

print("rp.orient()")

print(rp.orient())

#demo

#14 kernal names:identity edge-all edge-H edge-V sharp gauss-3 gauss-5 boxblur unsharp

#gradient-H gradient-V sobel-H sobel-V emboss

z=conv2d(c, "identity")

print("identity")

print(z)

z=conv2d(c, "edge-all")

print("edge-all")

print(z)

z=conv2d(c, "edge-H")

print("edge-H")

print(z)

z=conv2d(c, "edge-V")

print("edge-V")

print(z)

z=conv2d(c, "sharp")

print("sharp")

print(z)

z=conv2d(c, "gauss-3")

print("gauss-3")

print(z)

z=conv2d(c, "gauss-5")

print("gauss-5")

print(z)

z=conv2d(c, "boxblur")

print("boxblur")

print(z)

z=conv2d(c, "unsharp")

print("unsharp")

print(z)

z=conv2d(c, "gradient-H")

print("gradient-H")

print(z)

z=conv2d(c, "gradient-V")

print("gradient-V")

print(z)

z=conv2d(c, "sobel-H")

print("sobel-H")

print(z)

z=conv2d(c, "sobel-V")

print("sobel-V")

print(z)

z=conv2d(c, "emboss")

print("emboss")

print(z)

#testing end

d='C:\\Users\\peace\\Desktop\\12\\HGJ\\'

folders = list(filter(lambda x: os.path.isdir(os.path.join(d, x)), os.listdir(d)))

print(folders)

for f in range(0,len(folders)):

data\_path = d+str(folders[f])+'\\CT\\image\\'

rtss\_path=d+str(folders[f])+'\\CT\\RTS\\'

masks\_path=d+str(folders[f])+'\\CT\\masks\\'

overlayed\_path=d+str(folders[f])+'\\CT\\overlayed\\'

jpg\_path=d+str(folders[f])+'\\CT\\jpg\\'

masked\_path=d+str(folders[f])+'\\CT\\masked\\'

g = glob(data\_path + '/\*.dcm')

data\_dict = mat73.loadmat(rtss\_path+'RTSS.mat')

data = np.array(data\_dict)

a=data\_dict['contours']['Segmentation'].astype(np.uint64)

lst\_numpy\_arrs, ids = extract\_img\_array(data\_path, getID=True)

def load\_scan(path):

slices = [dicom.read\_file(path + '/' + s) for s in os.listdir(path)]

slices.sort(key = lambda x: int(x.InstanceNumber))

try:

slice\_thickness = np.abs(slices[0].ImagePositionPatient[2] - slices[1].ImagePositionPatient[2])

except:

slice\_thickness = np.abs(slices[0].SliceLocation - slices[1].SliceLocation)

for s in slices:

s.SliceThickness = slice\_thickness

return slices

def get\_pixels\_hu(scans):

image = np.stack([s.pixel\_array for s in scans])

image = image.astype(np.int16)

image[image == -2000] = 0

intercept = scans[0].RescaleIntercept

slope = scans[0].RescaleSlope

if slope != 1:

image = slope \* image.astype(np.float64)

image = image.astype(np.int16)

image += np.int16(intercept)

return np.array(image, dtype=np.int16)

ide=folders[f]

patient = load\_scan(data\_path)

imgs = get\_pixels\_hu(patient)

np.save('output\_csv\\'+"fullimages"+ide+".npy" , imgs)

file\_used=('output\_csv\\'+"fullimages"+ide+".npy")

n\_array = np.load(file\_used).astype(np.float64)

for i in range(0,len(g)):

j=a[i]

if sum(sum (j))!=0:

plt.imsave(masks\_path+str(i)+'mask.jpg',a[i])

plt.imsave(jpg\_path+str(i)+'image.jpg',n\_array[i])

c=a[i]\*n\_array[i]

plt.imsave(masked\_path+str(i)+'masked.jpg',c)

plt.imshow(c, interpolation='None', cmap='gray',alpha=0.5)

#plt.show()

'''

#features for all dicom images

dataframe = get\_df\_from\_img\_array(n\_array,ids, getId = True)

print("Extracted features for all images : ")

df = pd.DataFrame(dataframe)

df.to\_csv('1.csv')

'''

'''

#from dcm images features without id

dfa = get\_df\_from\_path(data\_path , [])

print("5 df from top are: ")

df = pd.DataFrame(dfa)

df.to\_csv('output\_csv\\'+str(folders[f])+'.csv')

'''

#features for masked images

X\_data = []

masked\_files = glob(masked\_path+'\*.jpg')

for myFile in masked\_files:

masked\_image = cv2.imread (myFile)

gray\_image = cv2.cvtColor(masked\_image, cv2.COLOR\_BGR2GRAY)

X\_data.append (gray\_image)

#print('X\_data shape:', np.array(X\_data).shape)

dataframe = get\_df\_from\_img\_array(X\_data,ids, getId = True)

df = pd.DataFrame(dataframe)

print("Extracted features for masked images for patient's ID : "+folders[f])

df.to\_csv('output\_csv\\'+str(folders[f])+'.csv')

'''

#features for overlayed images

masks\_files = glob(masks\_path+'\*.jpg')

jpg\_files = glob(jpg\_path+'\*.jpg')

for i in range(0,len(masked\_files)):

img=cv2.imread(masks\_files[i])

im2=cv2.imread(jpg\_files[i])

dst = cv2.addWeighted(img, 0.5, im2, 0.5, 0)

plt.imsave(overlayed\_path+str(i)+'overlayed.jpg',dst)

Y\_data = []

overlayed\_files=glob(overlayed\_path+'\*.jpg')

for myFile in overlayed\_files:

overlayed\_image = cv2.imread (myFile)

gray\_image = cv2.cvtColor(overlayed\_image, cv2.COLOR\_BGR2GRAY)

Y\_data.append (gray\_image)

#print('X\_data shape:', np.array(X\_data).shape)

dataframe = get\_df\_from\_img\_array(Y\_data,ids, getId = True)

print("Extracted features for overlayed images : ")

df = pd.DataFrame(dataframe)

df.to\_csv('output\_csv\\'+str(folders[f])+'.csv')

'''

final = 'C:\\Users\\peace\\Desktop\\12\\output\_csv\\\*.csv'

df = pd.concat((pd.read\_csv(f) for f in glob(final, recursive=True)), ignore\_index=True)

df.to\_csv('final.csv')

print("All csv files are combined as one. :)")