DETECTION AND PREDICTION OF LUNG CANCER CELLS

Presented by:
Rohan Tuli

Mentors:
Dr Dhirendra Mishra (Academic)
Dr Nagachetan Bangalore (External)

Radiological Assistance:
Dr Nilesh Haran

Contents

Introduction

Problem Definition

Scope

Literature Review

Architecture and Design

Implementation and Demo

Societal and practical implications

Conclusion and Future Scope

Introduction

Lung cancer is one of the most life-threatening cancers in the world. It is the leading cancer killer in both men and women worldwide and has one of the highest mortality rate among all other types of cancer.

One of the major reasons why this cancer is so detrimental is because it is extremely hard to diagnose in early stages.

It may take a couple of years for the malignant tumor to grow, and this disease does not exhibit any specific symptoms in early stages.

Our project is mainly concerned with the 'Analysis of Malignant Tumours for Lung Cancer patients', expediting the early detection of lung cancer being the objective which is to be achieved.

Problem Definition

Analysis of Malignant Tumours for Lung Cancer Patients

Objective: Detection of Malignant Tumour

We have used computer aided diagnosis techniques like image processing and machine learning for the development of an outline process for extraction and detection of lung tumours, thus supplementing and humanizing the physician's ability to spot abnormalities.

Scope

SCOPE DESCRIPTION

Explore existing strategies, research methodologies & limitations.

Augmenting current medical services for diagnosing & treating lung cancer.

The entire project revolves around the idea of detection of malignant tumour and the analysis of cancer progression for early diagnosis and prognosis of lung cancer for facilitating effective and accurate decision making.

Effective and early cancer prognosis after the cancer has been diagnosed promises an increase in the survival rate among the lung cancer patients.

We have explored lung CT scan DICOM images as our raw input data set. The necessary and relevant data pre-processing techniques have been applied on the data set in order to make it suitable for further processing.

Texture feature analysis has been carried out, as early evidence suggested that texture analysis has the potential to augment diagnosis and characterization as well as improve tumor staging and therapy response assessment in oncological practice.

We have automated the entire process under the expertise of a radiologist in order to deliver an efficient and robust model.

Literature Review

Makaju et al., 2018	This paper revolves around reviewing methods related to detection of cancerous nodules and proposing their own method in order to increase accuracy and efficiency.
Ganeshan & Miles, 2013	The authors are mainly focusing on heterogeneity and how it can be used to detect cancerous cells from biomarker images. The authors have shown that CT texture analysis has the potential to be a useful adjunct in clinical oncologic imaging, providing important information about tumour characterization, prognosis and treatment prediction and response.
Ng et al., 2013	According to this paper, it has been mentioned that tumour heterogeneity plays an important role in prognosis as high intra-tumoural heterogeneity may be associated with higher tumour grades. The authors have used CT (Computed Tomography) images to extract Entropy (E) and Uniformity (U).
Ganeshan et al., 2012	The authors objective was to establish the potential for tumour heterogeneity in non-small cell lung cancer (NSCLC) as assessed by CT texture analysis (CTTA) to provide an independent marker of survival for patients with NSCLC.

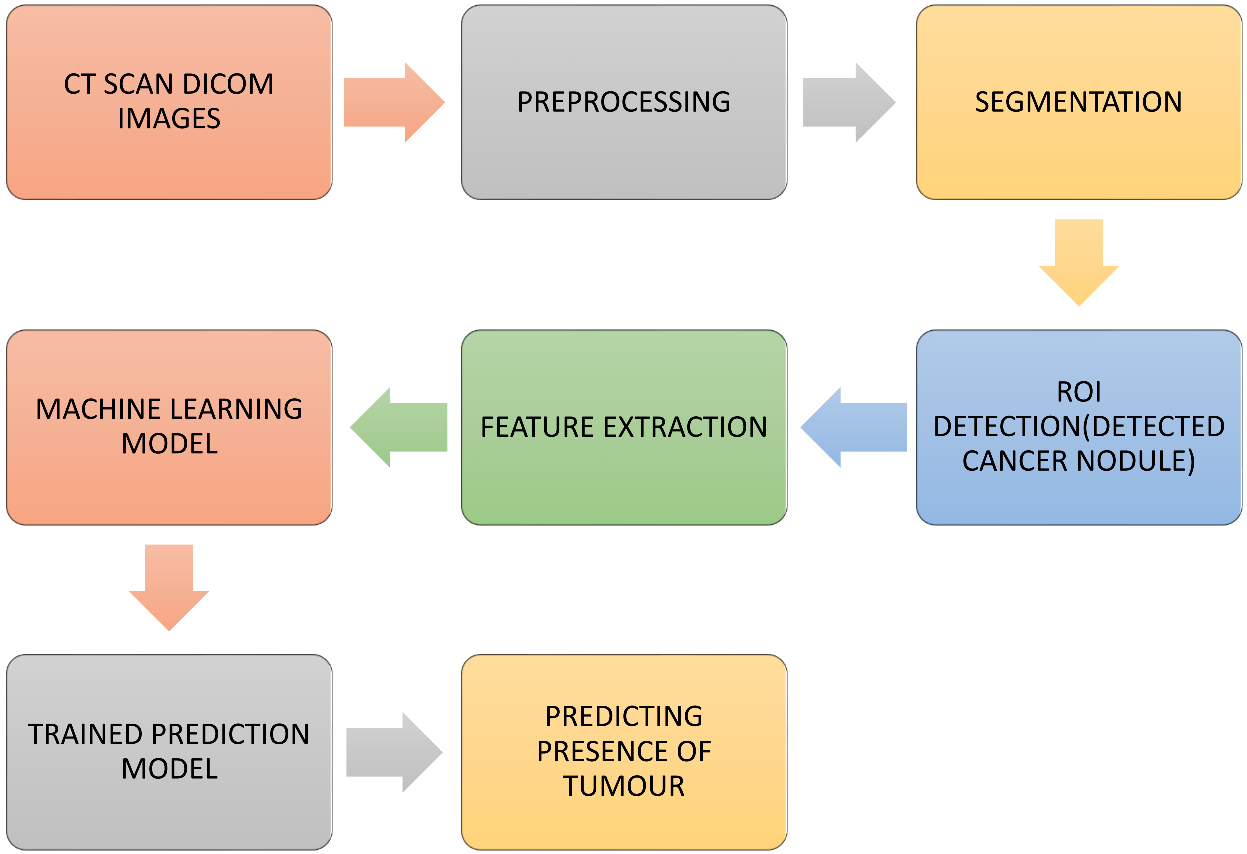
Literature Review

Dilpreet & Yadwinder, 2014	The authors have explored, discussed and listed advantages and disadvantages of several segmentation techniques. The authors have suggested that the methods discussed could be applied in various fields of medical imaging & satellite imaging.
Sharma et al., 2010	The author’s motive is to discuss the problems encountered in segmentation of CT images. And also implying the relative merits and limitations of methods currently available.
Planckt & Mueller, 1977	The authors presented methods for edge segmentation of satellite images. The authors used Sobel, Prewitt, Roberts, Canny & Laplacian operators. Prewitt was found to be best technique for edge detection.
Lin et al., 2000	The authors have discussed "The unseeded region growing (URG) algorithm" which is a derivative of seeded region

Literature Review

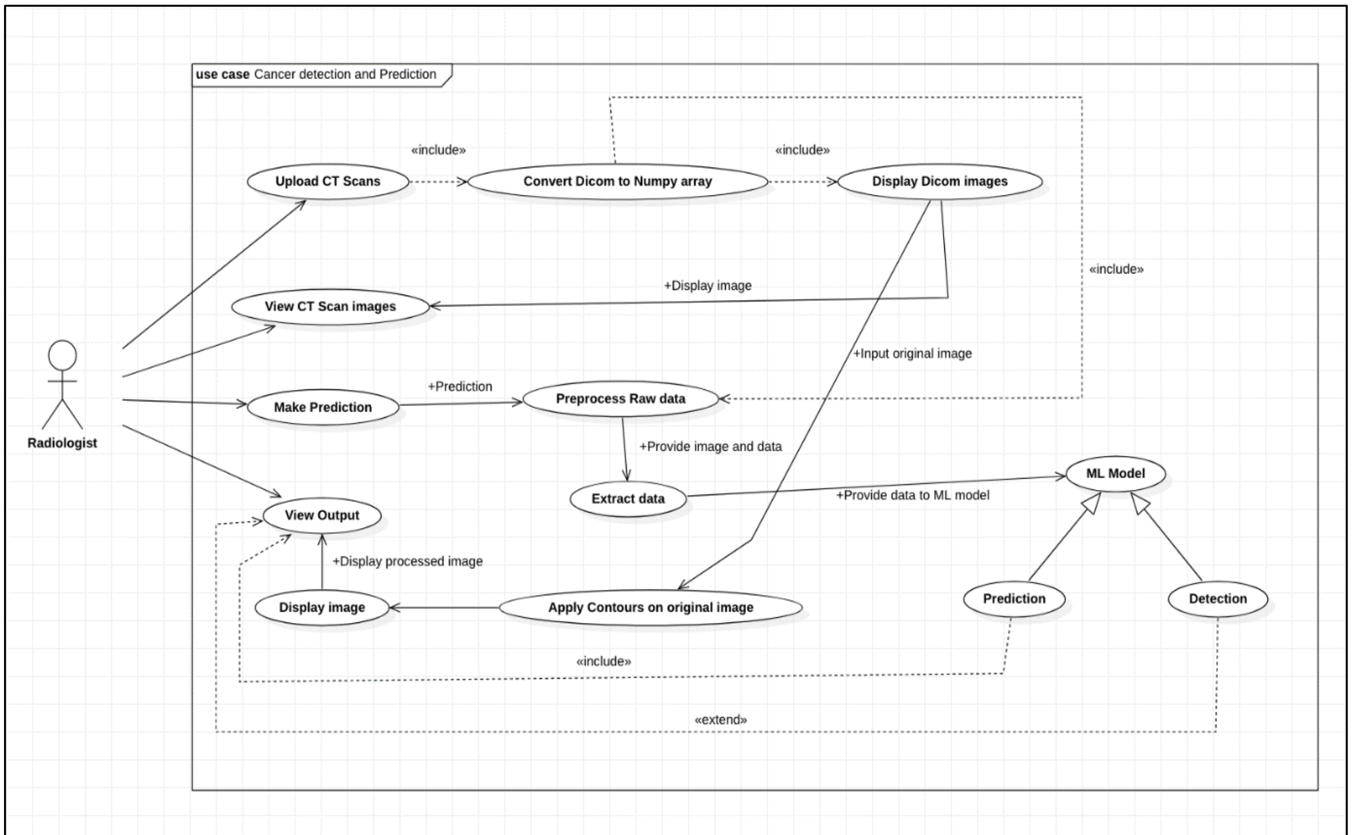
Wang, n.d.	The authors have elucidated that unlike region growing, fast scanning algorithm do not need seed point. The result is as good as that of region growing method but the running time is much less. The overall output is less optimized then region growing method but time taken is substantially less.
Mohanaiah et al., 2013	The authors have defined Gray Level Co-Occurrence Matrix (GLCM) method for extracting four statistical texture parameters. These features are useful in motion estimation of videos and in real time pattern recognition applications like Military & Medical Applications.
Zayed & Elnemr, 2015	The authors have presented an evaluation of the Haralick texture features that can be used in order to detect and differentiate abnormalities within the lungs for cancer. While the results given by the author are promising, there is further work that can be done in detecting abnormalities.
Kourou et al., 2015	The authors have discussed about ML methods for cancer prediction and prognosis. The authors suggest that the integration of multidimensional heterogeneous data combined with the application of different techniques for feature selection and classification can provide promising tools for inference in the cancer domain.

Architecture and Design



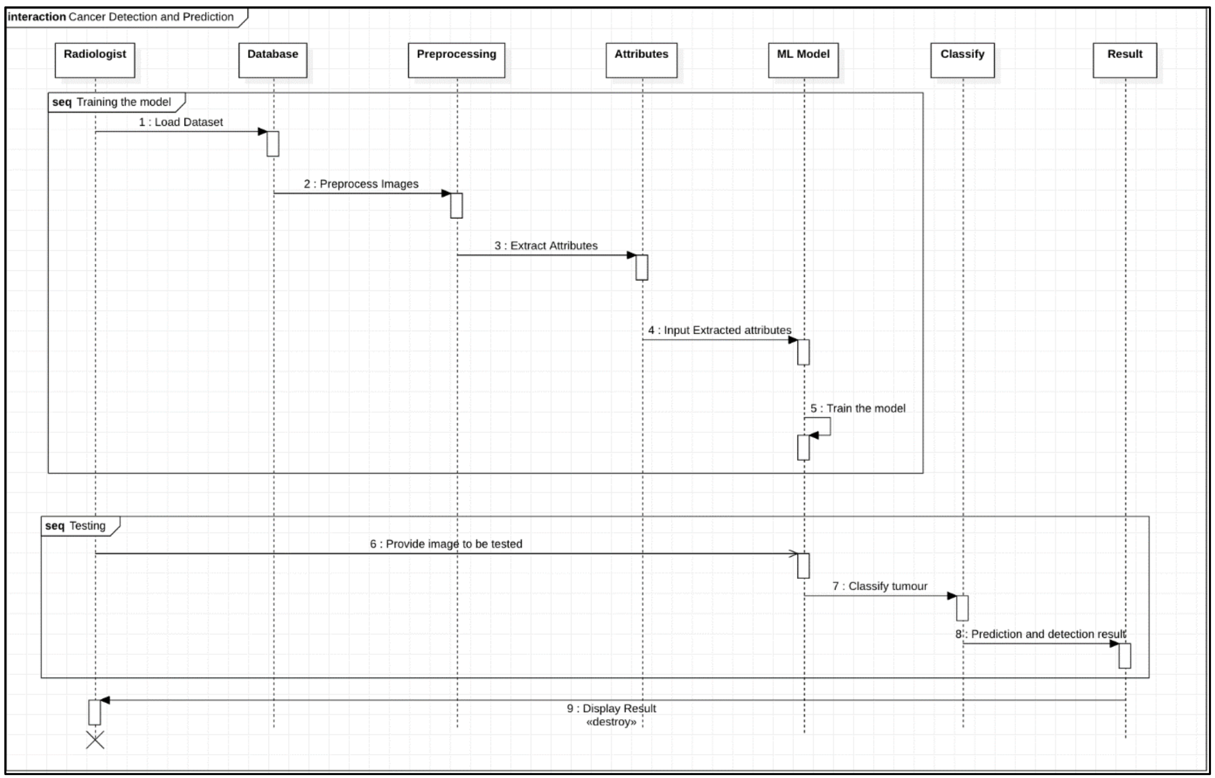
UML Diagrams

Use Case Diagram



UML Diagrams

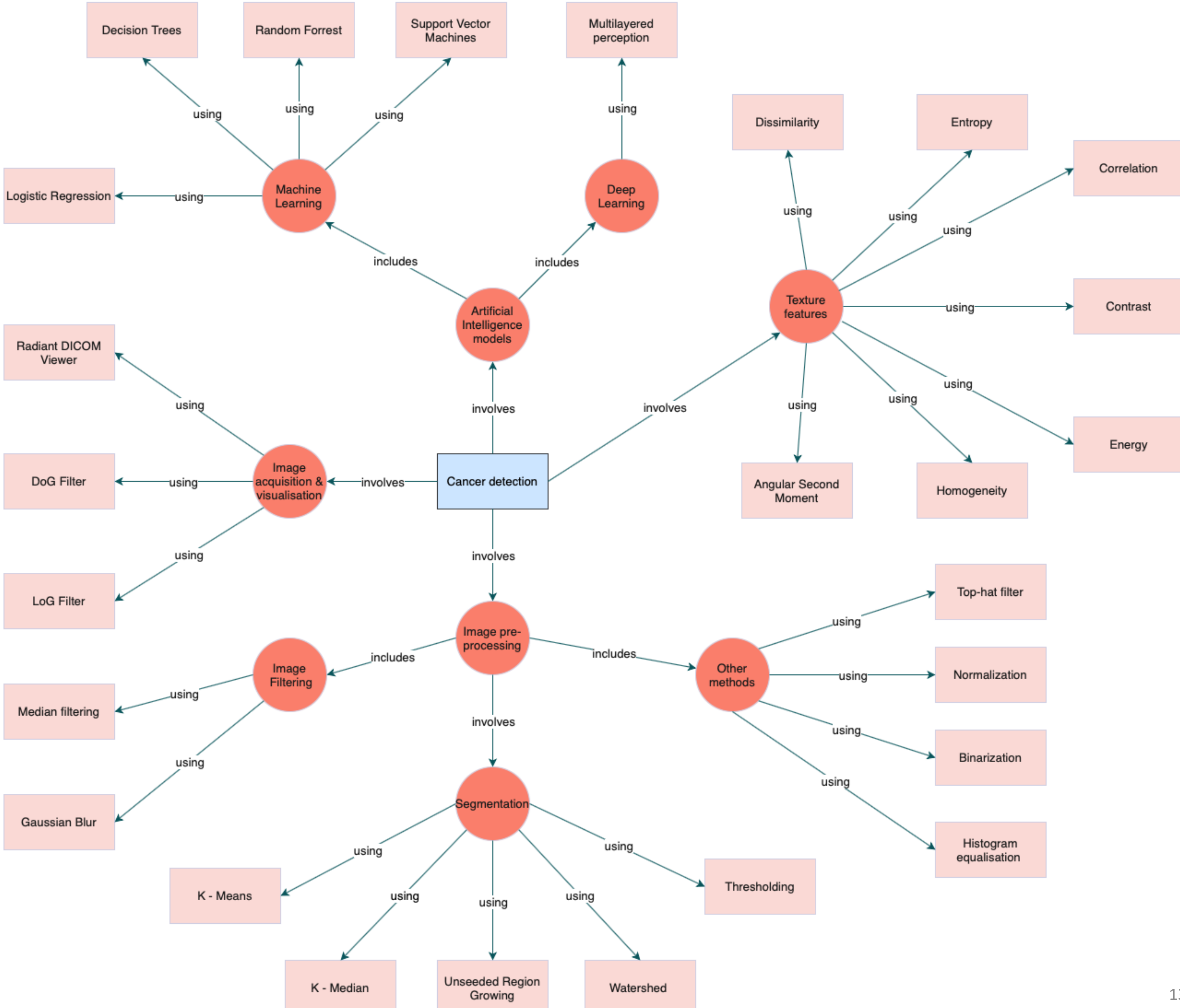
Sequence Diagram



Implementation

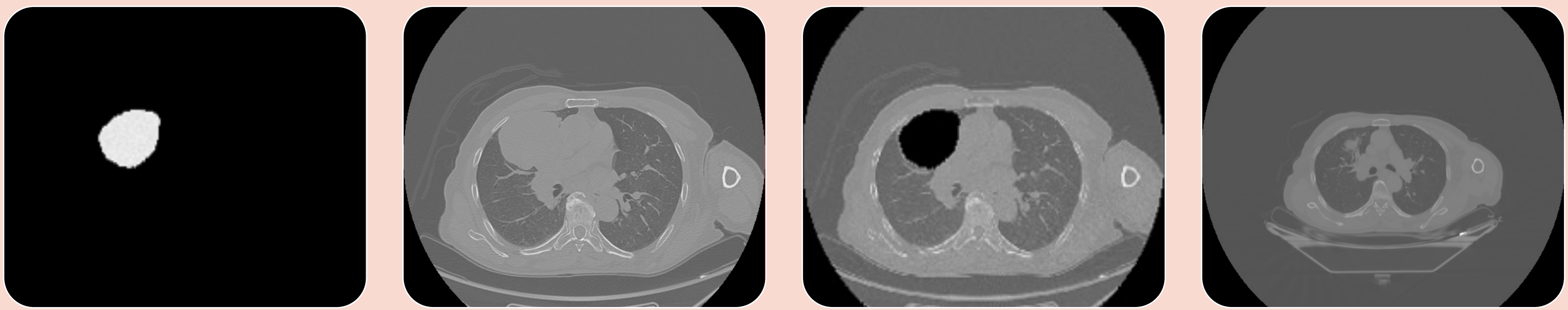
Modules Implemented

- Database Creation
- Preprocessing
- Segmentation
- Texture Feature Extraction
- Artificial Intelligence



Cancer detection
involves
Artificial Intelligence models
includes
Machine Learning
Deep Learning
using
Decision Trees
Random Forrest
Support Vector Machines
Logistic Regression
Multilayered perception
Texture features
Dissimilarity
Entropy
Correlation
Contrast
Energy
Homogeneity
Angular Second Moment
Image acquisition & visualisation
Radiant DICOM Viewer
DoG Filter
LoG Filter
Image pre-processing
Image Filtering
Median filtering
Gaussian Blur
Other methods
Top-hat filter
Normalization
Binarization
Histogram equalisation
Segmentation
K - Means
K - Median
Unseeded Region Growing
Watershed
Thresholding

Database Creation



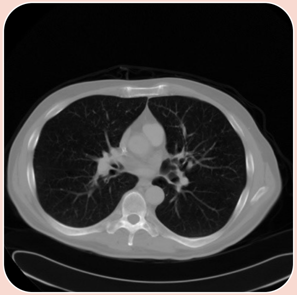
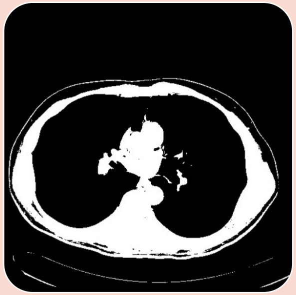
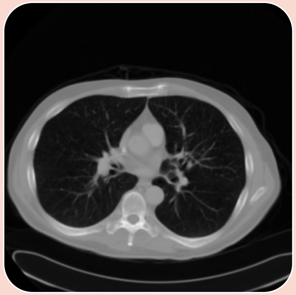
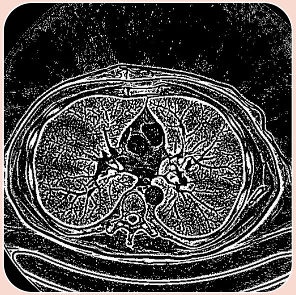
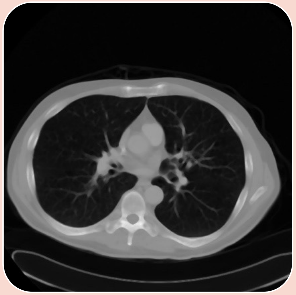
True Class

True True Class

False Class

False False Class

Preprocessing

Normalization

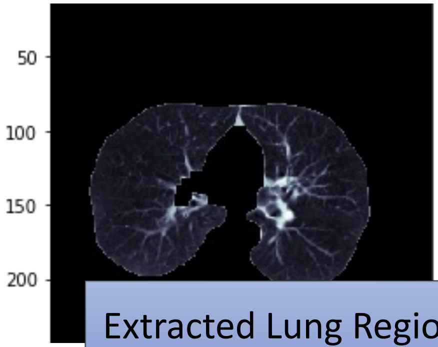
Binarization

Gaussian Blur

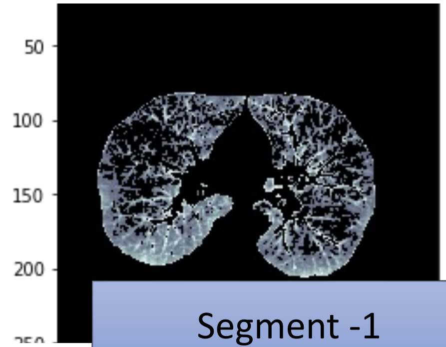
Difference of Gaussian

Median Blur

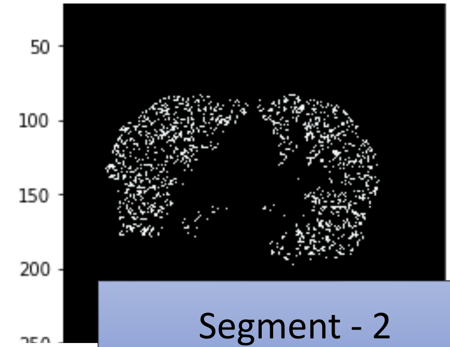
K - Means **Segmentation**



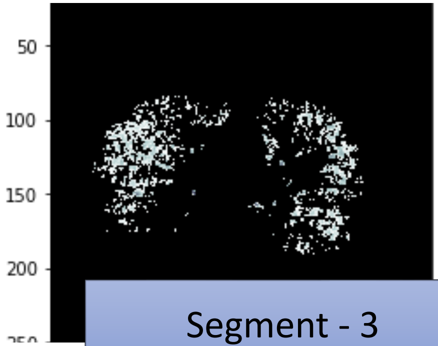
Extracted Lung Region



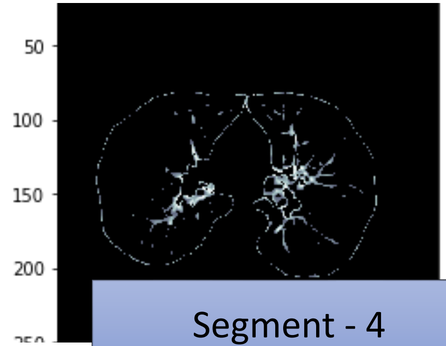
Segment -1



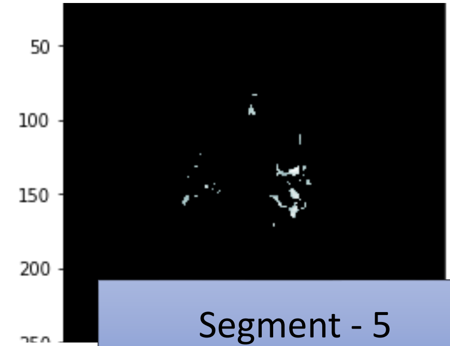
Segment - 2



Segment - 3

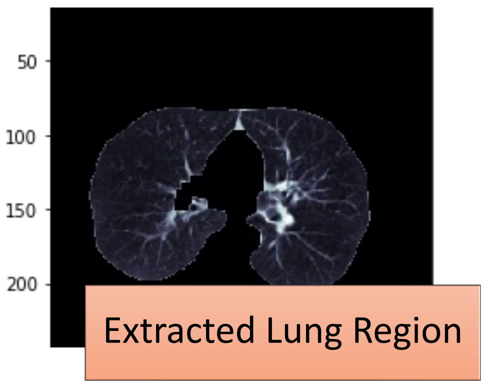


Segment - 4

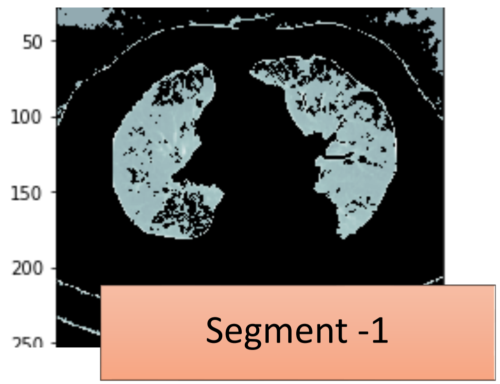


Segment - 5

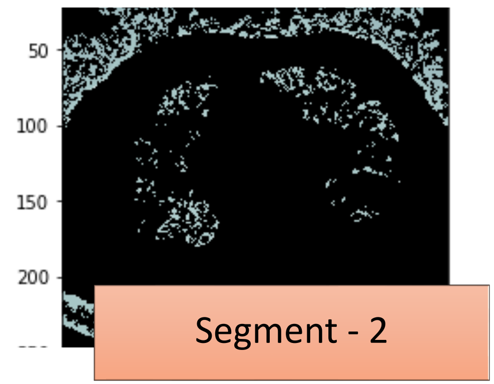
Unseeded Region Growing **Segmentation**



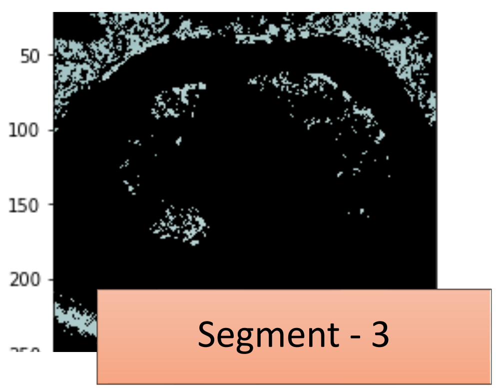
Extracted Lung Region



Segment -1



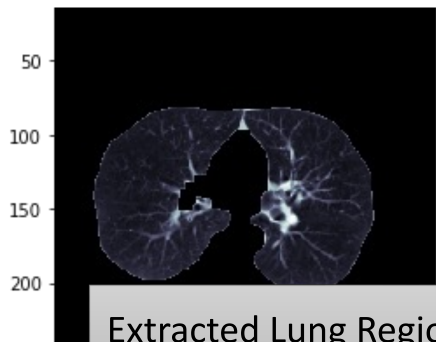
Segment - 2



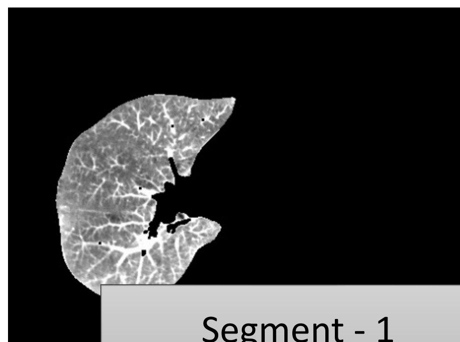
Segment - 3

Segment - 4

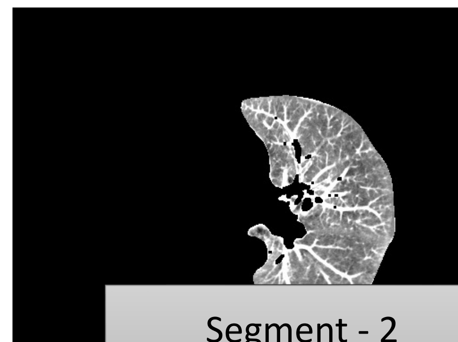
Watershed **Segmentation**



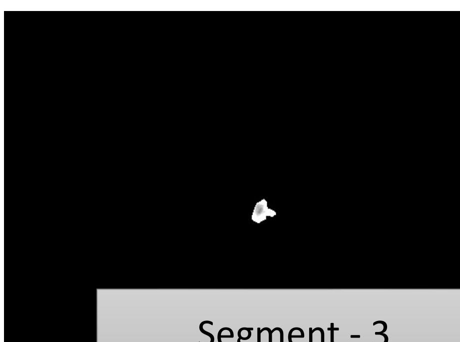
Extracted Lung Region



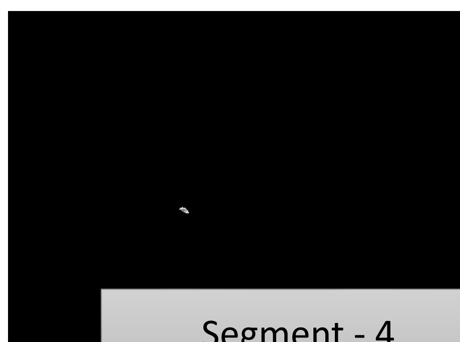
Segment - 1



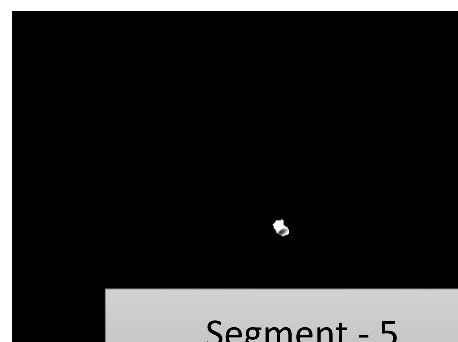
Segment - 2



Segment - 3

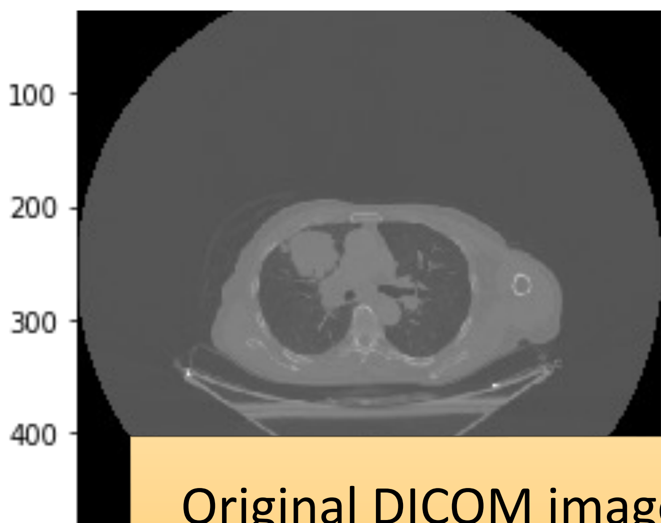


Segment - 4



Segment - 5

Segmentation by thresholding **Segmentation**



Original DICOM image

Segmented DICOM image

Texture Feature Extraction

Texture features are computed from the statistical distribution of observed combinations of intensities at specified positions relative to each other

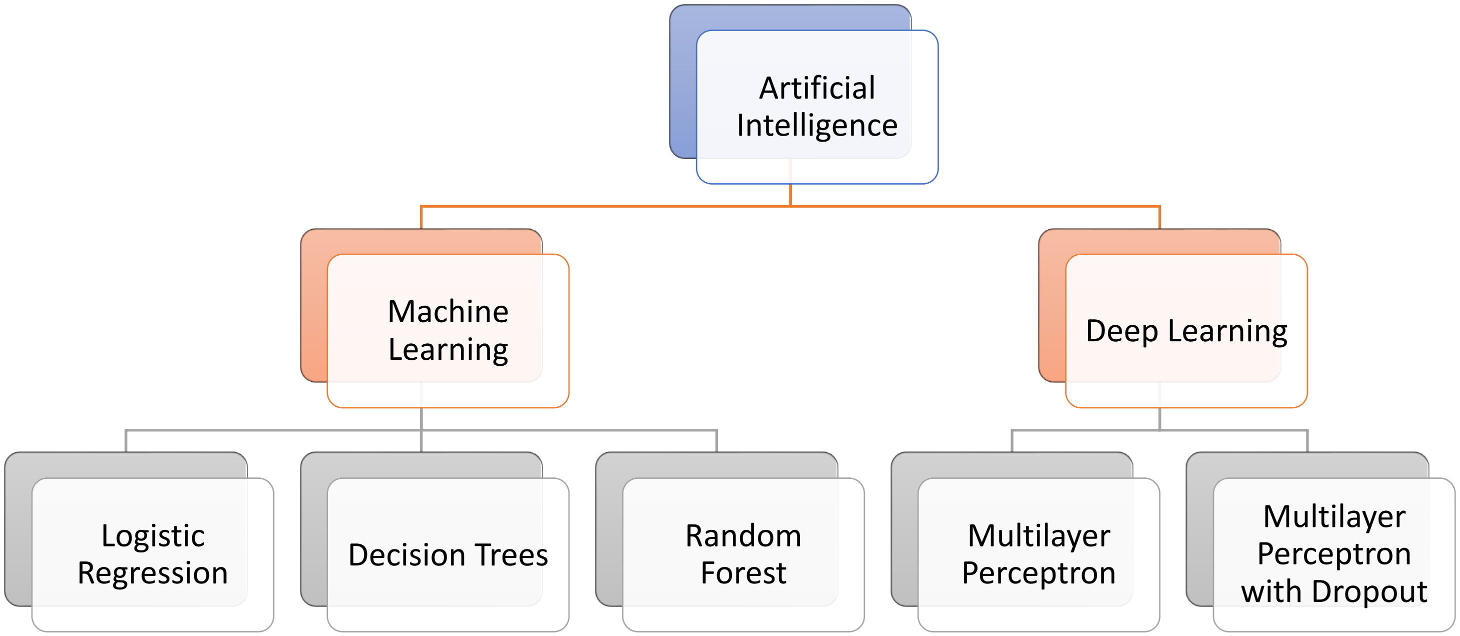
Gray level co-occurrence Matrix (GLCM)

- Method of extracting relevant second order statistical texture features
- Statistical features such as Contrast, Correlation, Homogeneity, Energy, Entropy, Dissimilarity & Angular Second Moment are calculated over a range of distances and angles

Customized GLCM

- Based on the concept of GLCM with the advantage of discarding undesirable background values
- Considers only the foreground for Texture Feature calculation

Artificial Intelligence Methodology



Machine Learning Methodology

Type	Model	Method	Confusion matrix				Precision	Recall	f1 score
			True positive	True negative	False positive	False negative			
Testing	Logistic Regression	-	869	610	88	176	0.84	0.85	0.85
	Decision Trees	Entropy	1036	685	13	9	0.99	0.99	0.99
		Gini Index	1026	689	9	19	0.98	0.98	0.98
	Random Forest	Entropy	1029	669	29	16	0.97	0.97	0.97
		Gini Index	1030	660	38	15	0.97	0.97	0.97

Legends	
Colour	**Hierarchy**
Blue	Highest accuracy
Yellow	Lowest accuracy

Summary of Machine Learning Models

Deep Learning Methodology

Type	Model	Number of hidden	Number of neurons in hidden layer				Model		Batch size	Confusion matrix				Precision	Recall	f1 score
			h1	h2	h3	h4	Accuracy	Loss		True positive	True negative	False positive	False negative			
Testing	1	1	8	-	-	-	0.844	-	-	792	545	247	0	0.88	0.84	0.84
	2	1	6	-	-	-	0.9311	0.1502	5	766	451	72	18	0.94	0.92	0.93
	2	1	6	-	-	-	0.9327	0.1436	12	760	459	64	24	0.94	0.92	0.93
	3	1	128	-	-	-	0.9273	0.1386	5	746	466	57	38	0.93	0.92	0.92
	3	1	128	-	-	-	0.9396	0.1231	12	764	464	59	20	0.94	0.93	0.94
	4	2	128	256	-	-	0.9265	0.1352	5	763	448	75	21	0.93	0.91	0.92
	4	2	128	256	-	-	0.9503	0.1052	12	766	476	47	18	0.95	0.95	0.95
	5	2	256	512	-	-	0.9288	0.1434	5	774	440	83	10	0.94	0.91	0.92
	5	2	256	512	-	-	0.9526	0.0984	12	761	484	39	23	0.95	0.95	0.95
	6	3	256	512	1024	-	0.899	0.2062	5	719	456	67	65	0.89	0.89	0.89
	6	3	256	512	1024	-	0.9434	0.1025	12	759	474	49	25	0.94	0.94	0.94
	7	4	128	512	1024	2048	0.9036	0.161	5	784	397	126	0	0.93	0.88	0.89
	7	4	128	512	1024	2048	0.9166	0.1413	12	725	473	50	59	0.91	0.91	0.91

Legends	
Colour	Hierarchy
Blue	Highest accuracy
Yellow	Lowest accuracy
Green	Notable values

Summary of Deep Learning Models
(Models 2 - 7 have been implemented using Tensorflow)

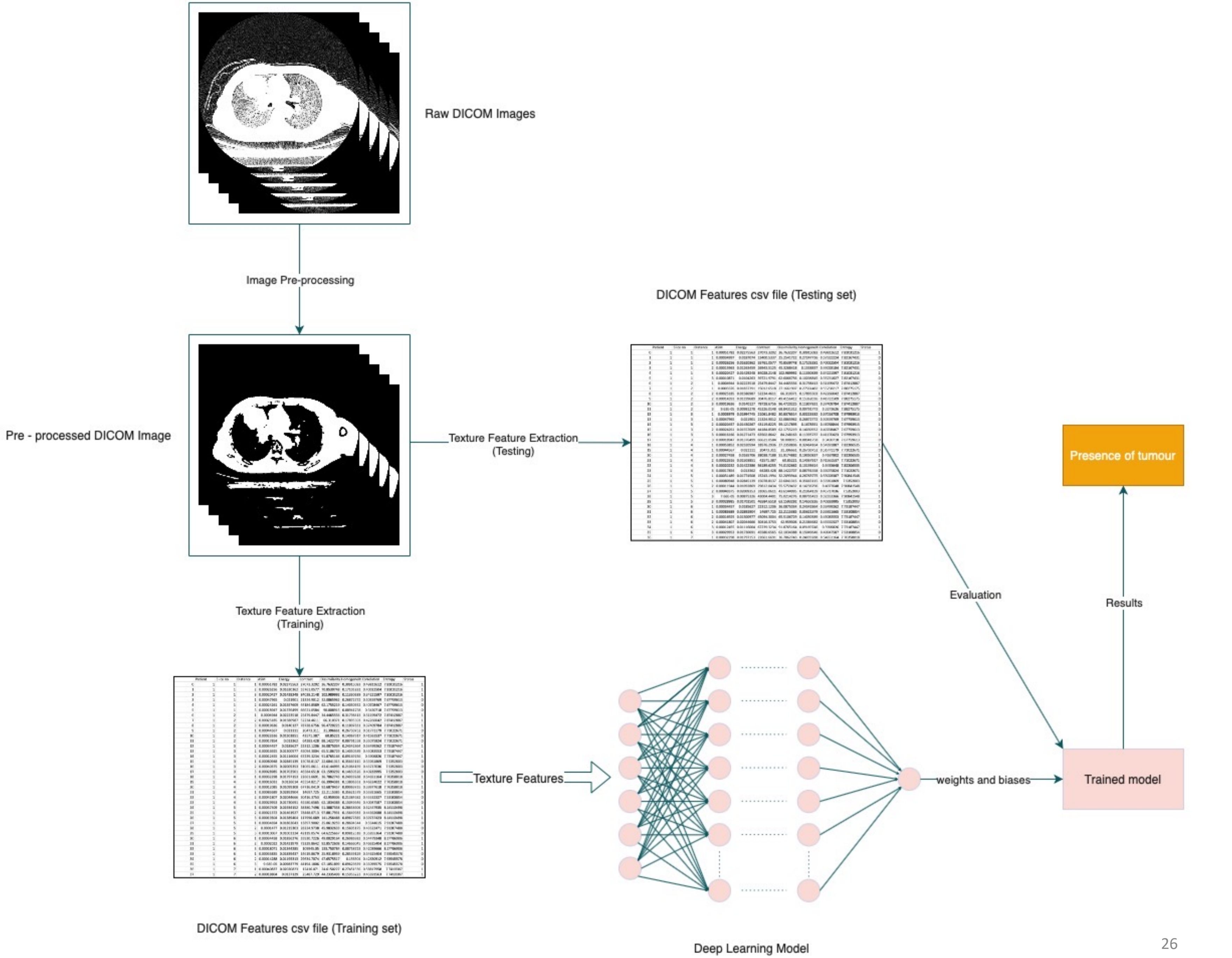
Deep Learning Methodology

Type	Model	Number of hidden	Number of dropout	Number of neurons in hidden layer				Model		Batch size	Confusion matrix				Precision	Recall	f1 score
				h1	h2	h3	h4	Accuracy	Loss		True positive	True negative	False positive	False negative			
Testing	1	1	2	6	-	-	-	0.7324	0.5735	6	82	74	29	28	0.73	0.73	0.73
				6	-	-	-	0.77	0.549	2	84	80	23	26	0.77	0.77	0.77
	2	1	2	128	-	-	-	0.7277	0.5329	6	86	69	34	24	0.73	0.73	0.73
				128	-	-	-	0.6995	0.5518	2	83	66	37	27	0.7	0.7	0.7
	3	2	3	128	256	-	-	0.7371	0.5441	6	79	78	25	31	0.74	0.74	0.74
				128	256	-	-	0.6854	0.6011	2	81	65	38	29	0.69	0.68	0.68
	4	2	3	256	512	-	-	0.831	0.4123	6	84	93	10	26	0.84	0.83	0.83
				256	512	-	-	0.8357	0.4287	2	92	86	17	18	0.84	0.84	0.84
	5	3	4	256	512	1024	-	0.7136	0.5463	6	69	83	20	41	0.72	0.72	0.71
				256	512	1024	-	0.6948	0.584	2	90	58	45	45	0.71	0.69	0.69
	6	4	5	128	512	1024	2048	0.7371	0.5466	6	90	67	36	20	0.74	0.73	0.73
				128	512	1024	2048	0.7324	0.5793	2	83	73	30	27	0.73	0.73	0.73

Legends	
Colour	Hierarchy
Blue	Highest accuracy
Yellow	Lowest accuracy
Green	Notable values

Summary of Deep Learning Models with Dropout (p = 0.5)
(Models 2 - 7 have been implemented using Tensorflow)

DEMONSTRATION


Raw DICOM Images
Image Pre-processing
Pre - processed DICOM Image
Texture Feature Extraction
(Testing)
DICOM Features csv file (Testing set)
Texture Feature Extraction
(Training)
DICOM Features csv file (Training set)
Texture Features
Deep Learning Model
weights and biases
Evaluation
Trained model
Results
Presence of tumour

Project Highlights

Automated lung cancer detection model has been developed which involves minimal human interaction

Extensive research has been conducted into unifying a single work flow, incorporating image processing, texture feature analysis and artificial intelligence

A custom module has been developed for texture feature extraction based on the Gray-Level Co-Occurrence Matrix (GLCM) texture feature extraction methodology in which only the foreground has been considered for the retrieval of statistical texture features

The prediction results of the developed model have been verified by Dr. Nilesh Haran (MBBS, MD in Radio-Oncology)

Societal and Practical Implications

Our objective is to develop a cost efficient and accurate pipeline of early detection of cancerous tumour.

We aim at incorporating texture analysis using an AI model into routine clinical workflow.

A well-trained AI model will result in an efficient and robust model. Eliminating human error and assisting the appropriate person.

Conclusion and Future Scope

We have found CT Texture Analysis (CTTA) methodology promising, which makes use of Haralick texture features to provide crucial statistical information

Although several classifiers were created, only a few of them gave us usable results thus, motivating us to explore further concepts which are robust to outliers and efficient at detecting anomalies leading us to achieve near ideal results

Devise and implement an algorithm which successfully detects the spread, diagnosis the stage and predicts the direction and rate of the cancer spread

Extend our research for detecting cancerous tumors in other organs such as breast, brain, etc

Q&A

THANK YOU!