3-D IMAGE ANALYSIS TECHNIQUES FOR LUNG CANCER DETECTION ON COMPUTED TOMOGRAPHY (CT) SCANS

HINA SHAKIR

02-281141-001

A thesis submitted in fulfillment of the

requirements for the award of the degree of

Doctor of Philosophy (Electrical Engineering)

Department of Electrical Engineering

BAHRIA UNIVERSITY, KARACHI

DECEMBER, 2021

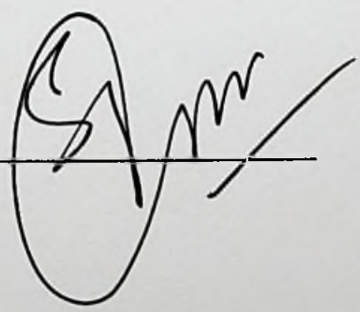
Approval for Examination

Scholar's Name: Hina Shakir Registration No: 37188

Programme of Study: PhD

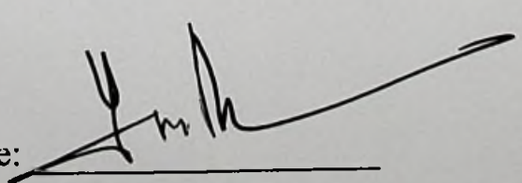
Thesis Title: 3-D Image Analysis Techniques for Lung Cancer Detection on Computed Tomography (CT) Scans

It is to certify that the above scholar's thesis has been completed to my satisfaction and, to my belief, its standard is appropriate for submission for examination. I have also conducted plagiarism test of this thesis using HEC prescribed software and found similarity index 15% that is within the permissible limit set by the HEC for the PhD degree thesis. I have also found the thesis in a format recognized by the BU for the PhD thesis.

Principal Supervisor's Signature: 

Date: 08/12/21

Name: Sr. Prof. Dr. Haroon Rasheed

Co- Supervisor's Signature: 

Date: 08/12/21

Name: Dr. Tariq Mairaj Rasool Khan

Author's Declaration

I, Hina Shakir hereby state that my PhD thesis titled 3-D Image Analysis Techniques for Lung Cancer Detection on Computed Tomography (CT) Scans is my own work and has not been submitted previously by me for taking any degree from this university Bahria University rule or anywhere else in the country/world.

At any time if my statement is found to be incorrect even after my graduation, the University has the right to withdraw/cancel my PhD degree.

9/12/21

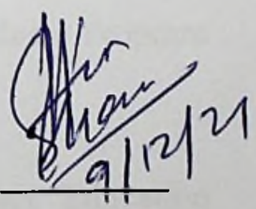
Name of scholar: Hina Shakir

Date: 6/12/2021

Plagiarism Undertaking

I, solemnly declare that research work presented in the thesis titled 3-D Image Analysis Techniques for Lung Cancer Detection on Computed Tomography (CT) Scans is solely my research work with no significant contribution from any other person. Small contribution / help wherever taken has been duly acknowledged and that complete thesis has been written by me.

I understand the zero tolerance policy of the HEC and Bahria University towards plagiarism. Therefore, I as an Author of the above titled thesis declare that no portion of my thesis has been plagiarized and any material used as reference is properly referred / cited. I undertake that if I am found guilty of any formal plagiarism in the above titled thesis even after award of PhD degree, the university reserves the right to withdraw / revoke my PhD degree and that HEC and the University has the right to publish my name on the HEC / University website on which names of scholars are placed who submitted plagiarized thesis.

Scholar/Author's Sign: 

Name of the Scholar: Hina Shakir

Acknowledgment

I would like to express my sincere gratitude to my supervisor Prof. Dr. Haroon Rasheed and my co-supervisor Dr. Tariq Mairaj Rasool Khan for their constant help, invaluable suggestions and encouragement that helped me in all the times to pursue this PhD dissertation work.

I am grateful to my foreign examiners, Prof. Dr. Valentina Emilia Balas, Aurel Vlaicu University of Arad, Romania and Dr. Jeffrey Will, Valparaiso University, USA for their expert advice to improve my thesis. I am also grateful to my advisory committee members Dr. Shahid Sheikh, Dr. Farah Haroon, Dr. Anzar Alam and Dr. Ghazanfar Monir for their valuable insights and useful advices during entire my PhD program. The guidance provided by my internal defense examination committee members including Dr. Najeed Ahmed Khan, Dr. Imran Naseem, Dr. Yawar and Dr. Samabia Tehseen immensely helped me through my dissertation for which i am truly thankful to them.

Many thanks to the administrative staff at Electrical Engineering program especially the Head of the Department, Dr. Najam Amin and Post Graduate Program (PGP) coordinator, Dr. Mukesh Maheshwari for their timely assistance.

I would want to gratefully acknowledge the role of Higher Education Commission (HEC) for awarding research scholarship of 6 months visit to Michigan State University, USA. The foreign visit provided me with an excellent research exposure to carry out significant part of my research work successfully.

Special thanks to Professor Yiming Deng for the opportunity to join his research group at Michigan State University, USA. The insightful discussions i had with him and his interest in my research work truly inspired me to make progress in my research endeavors.

I am also obliged to my class mates and my friends for their companionship and lively conversations which elated me when my research work was not progressing as I expected.

I owe my deepest regards and thanks to my late father and my brother for being the pillars of support and understanding the importance of this journey for my professional and personal growth. I cannot thank my sister enough whose encouragement and love motivated me to continue working me towards my research pursuits.

ABSTRACT

Compelling clinical studies have shown that lung cancer screening allows for early diagnosis and treatment of lung cancer. Despite the high quality CT images available to the radiologists for decision making, missing lung cancer on the chest radiograph attributes to errors made by the radiologists in determining whether a lung nodule is malignant or not, on the basis of medical imaging and other clinical information. Several research studies have recommended Computer Aided Diagnosis (CADx) systems which facilitate early detection of lung cancer nodules at more curable stages. The significant and on-going research areas in CADx systems include automated segmentation systems and classification models for nodules characterization. The non-availability of large databases has been a limitation for the training and development of accurate diagnostic models.

The thesis aims to develop image analysis based methods in 3-D to improve the performance of CAD systems for lung cancer detection and pathological stage estimation. The contributed research work was carried out in the areas of segmentation and cancer diagnostic models respectively which are integral parts of a CAD system. During the first phase of research, a semi-automatic 3-D segmentation algorithm was developed for varying nodules sizes and shapes which enables accurate volumetric assessment and further analysis for nodules characterization in Computed Tomography (CT) images. The proposed model performs nodule segmentation by incorporating the geometric and region based properties of a lung nodule in CT images. An adaptive approach was developed to compute the mean gray level intensity of a lung nodule which is used as a stopping function besides the gradient properties to enable accurate nodule extraction from the CT images. The algorithm was tested and validated on multiple databases including LIDC, FDA, RIDER and CUMC with a mean spatial overlap of 85%.

As second research contribution, a Bayesian Inversion framework was devised to predict the cancer occurrence as well as its pathological stage. The proposed stochastic model was tested on NLST database, a repository of low dose longitudinal CT data sets and is aimed to address the constraint of limited number of annotated CT datasets in lung cancer for multi-class classification. To develop the proposed framework, quantitative imaging features known as radiomic features were extracted from the segmented nodules (benign and malignant), analyzed and ranked using an intuitive approach to identify the diagnostic features for cancer and its stage estimation respectively. The carried out research work showed that Surface Area and Sum Entropy are the two radiomic features that possess general phenotype/characteristics to differentiate between benign and cancerous tumors of lung, colon and head and neck cancer respectively. These highly discriminating features were validated by integration into non-linear mathematical equation and achieved an accuracy of 97.0% for lung nodule classification.

Another group of radiomic features was investigated in malignant nodules which can differentiate between early stage (stage I and stage II) and advanced stage (stage III and stage IV) cancer. Based on the analysis performed using regularization method and supervised feature selection, radiomic features which showed strong discrimination between early and advanced stage cancer were Sphericity, Large Dependence High Gray Level Emphasis (LDHGLE), Cluster Prominence, Small Area Emphasis and Strength respectively. These features were integrated into a mathematical equation for validation and achieved an accuracy of 86.84% when tested on data sets of malignant nodules with early and advanced stage cancer.

By incorporating the aforementioned formulated equations for cancer and its stage detection respectively, the proposed stochastic framework predicted the cancer occurrence as well as its pathological stage with an accuracy of 86%.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

III A

LIST OF ABBREVIATIONS

2-D	Two dimensional
3-D	Three Dimensional
ACM	Active Contour Model
ANOVA	Analysis of Variance
AOS	Additive Operator Splitting
AUC	Area under the Curve
CADx/CAD	Computer Aided Diagnosis
CCC	Concordance Correlation Coefficient
CI	Confidence Interval
CP	Cluster Prominence
CT	Computed Tomography
CTC	CT Colongraphy
CUMC	Columbia University Medical Center
CV	Coefficient of Variation
df	Degree of freedom
DC	Dice Coefficient
DCNN	Deep convolutional neural network

DenseBTNet	Dense Convolutional Binary-Tree Network
DWT	Discrete Wavelet Transform
$E(x)$	Expected/mean value of x
ESS	Effective Sample Size
FDA	Food and Drug Administration
FN	False Negative
fNCA	Feature based Neighborhood Component Analysis
FN	False Negative
FP	False Positive
FPR	False Positive Rate
GAC	Geodesic Active Contour
GGO	Ground Glass Opacity
GHz	Gigahertz
GLCM	Gray Level Co-Occurrence Matrix
GLDM	Gray Level Difference Method
GLMR	Generalized Linear Regression Model
GLRLM	Gray Level Run Length Matrix
GLSZM	Gray Level Size Zone Matrix
HNSCC	Head-and-Neck Squamous Cell Carcinoma
HU	Hounsfield Unit
IDRI	Image Database Resource Initiative
ILFS	Infinite Latent Feature Selection

IQR	Inter Quartile Range
LALGLE	Large Area Low Gray Level Emphasis
LASSO	Least Absolute Shrinkage and Selection Operator
LDA	Linear Discriminant Analysis
LDCT	Low Dose Computed Tomography
LDHGLE	Large Dependence High Gray Level Emphasis
LIDC	Lung Imaging Database Consortium
LLC	Local Learning-Based Clustering
MAP	Maximum a posteriori
MDCT	Multidetector CT
mm	milli meter
ml	milli litre
MLF	Mathematical likelihood function
MMR	Multiple Mismatch Repair
MRI	Magnetic Resonance Imaging
MR-T1	Magnetic Resonance Basic Pulse
mSv	millisievert
MVCNN	Multi-view Convolutional Neural Networks
NGTDM	Neighborhood Gray-Tone Difference Matrix
NLST	National Lung Screening Trial
NN	Neural Network
PC	Principal Component
PCA	Principal Component Analysis

PDE	Partial Differentiation Equation
PDF	Probability Density Function
PET	Positron Emission Tomography
PME	Posterior Mean Estimate
RADS	Reporting and Data System
RAM	Random Access Memory
RC	Repeatability Coefficient
RECIST	Response Evaluation Criteria In Solid Tumors
RIDER	Reference Image Database to evaluate Response Therapy
RMSE	Root Mean Square Error
RNN	Regression Neural Network
ROI	Region of Interest
ROC	Receiver Operating Characteristic
SDF	Signed Distance Function
SDHGLE	Small Dependence High Gray Level Emphasis
SDLGLE	Small Dependence Low Gray Level Emphasis
SE	Standard error
SFS	Sequential Forward Selection
SRLGLE	Short Run Low Gray Level Emphasis
SU	Stanford University
SVM	Support Vector Machine
SVR	Surface Volume Ratio
s_w^2	within-subject variance

T	Tesla
TCIA	The Cancer Imaging Archive
TP	True Positive
TPR	True Positive Rate
TN	True Negative
VIBE	Volumetric Interpolated Breath-hold Sequence
VOI	Volume of Interest
VQ	Vector Quantization
wCV	Within-subject Coefficient of Variance
ZV	Zone variance

LIST OF SYMBOLS

$\emptyset$	Level-set function
K	Curvature of the evolving curve
β	Function computed on level set function
div	Divergence operator
μl	Microlitre
μ	Computed mean
ε	Energy functional
α	Positive constant in heat equation
Ω	Open set in Real numbers domain
v	Lagrange multiplier
Δ	Step size
∇	Gradient operator
χ	Chi-square distribution
δ	Delta function
ξ	Convergence threshold
p	Scalar to control variability in probability distribution
λ	Negative likelihood ratio

$\Re^2$	2-D domain in Real numbers
σ	Variance
K	Conductance Parameter
γ	Damping coefficient
$\vec{N}$	Inward normal vector
∂	Partial derivative
Π	Product operator
$\sum$	Summation operator
$\|\|$	Absolute value
$\propto$	Proportionality symbol
$\approx$	Approximation symbol
$*$	Convolution operator
$\cup$	Union
$\cap$	Intersection