UNIVERSITI TEKNOLOGI MARA

EARLY-STAGE LUNG CANCER DIAGNOSIS USING NEW REGRESSION FEATURES AND MACHINE LEARNING

NURUL NAJIHA BINTI JAFERY

Thesis submitted in fulfilment
of the requirements for the degree of
Doctor of Philosophy
(Electrical Engineering)

Faculty of Electrical Engineering

September 2025

ABSTRACT

Lung cancer is the most common cancer worldwide and one of the leading causes of cancer-related deaths due to late detection. Radiologists typically diagnose lung cancer through the visual analysis of computed tomography (CT) scan images, a process that is tedious, time-consuming, and prone to errors. Additionally, variations in CT scan image intensity and the potential for misinterpretation of anatomical structures make it challenging to accurately identify cancerous cells. The TNM (Tumour, Node, Metastases) staging system is commonly used by doctors and radiologists to classify lung cancer progression. Early detection of lung cancer, particularly in the T1 and T2 stages, significantly improves survival rates, highlighting the importance of timely and accurate diagnosis. This study aims to develop an automated early-stage lung cancer diagnosis system using a new regression feature extraction method and machine learning techniques. The system is designed to assist radiologists and medical experts in diagnosing lung cancer and making treatment decisions. The methodology is divided into five stages: image acquisition, pre-processing, lung lesion detection, early-stage lung cancer diagnosis, and performance evaluation. The lung CT scan images used in this study were obtained from the Advanced Medical and Dental Institute (AMDI), Universiti Sains Malaysia (USM). In the pre-processing stage, a new segmentation method using geometrical features was proposed to segment lung lesion and non-lesion regions. For lung lesion detection, a new Regression Features (RFE) was introduced, generating four feature sets: RFE_1, RFE_2, RFE_3, and RFE_4. The best-performing set, RFE_2, was then fed into two proposed hybrid deep neural networks: Hybrid 1D-CNN-LSTM and VGG16-1D-LSTM, to classify lung lesion and non-lesion regions. Both models achieved an accuracy of 96%, with the Hybrid 1D-CNN-LSTM outperforming VGG16-1D-LSTM in AUC (0.91 vs. 0.81). Identified lung lesions were further analysed in the early-stage lung cancer diagnosis stage using machine learning classifiers, including Support Vector Machine (SVM), Gradient Boosting, AdaBoost, and Random Forest. Among these, Random Forest demonstrated the highest capability for automatically diagnosing early-stage lung cancer, achieving a cross-validated accuracy of 97.14% and an AUC of 0.9884. In the performance evaluation stage, the results were correlated with patient radiology reports to assess clinical relevance. The findings suggest that the proposed system has the potential to serve as an effective decision-support tool for radiologists in diagnosing early-stage lung cancer, ultimately improving early detection, patient outcomes, and clinical workflow efficiency.

ACKNOWLEDGEMENT

In the name of Allah S.W.T, the Most Gracious, the Most Merciful. I am deeply grateful for His blessings, which have given me the opportunity to embark on my PhD and successfully complete this long and challenging journey. My deepest gratitude and thanks go to my supervisor, Assoc. Prof. Ir. Ts. Dr. Hajah Siti Noraini Sulaiman, and my co-supervisors, Assoc. Prof. Dr. Muhammad Khusairi Osman, Assoc. Prof. Dr. Noor Khairiah A. Karim, and Assoc. Prof. Ir. Ts. Dr. Zainal Hisham Che Soh. Thank you for your continuous support, patience, and invaluable guidance in assisting me throughout this project. Your expertise, availability, and encouragement have been critical to the success of this research. I sincerely appreciate your willingness to share both my moments of joy and my moments of frustration.

I would also like to express my gratitude to the staff of the Advanced Medical and Dental Institute (AMDI), especially Suzana Ismail, for providing the necessary facilities, knowledge, and assistance. My appreciation extends to all the supportive staff from the Advanced Control System and Computing Research Group (ACSCRG) and the Centre for Electrical Engineering Studies, UiTM Pulau Pinang, whose contributions, both direct and indirect, have been instrumental in completing this project. Without their cooperation and assistance, this work would not have been possible.

Thanks to my beloved mother, , my beloved father, Jafery bin Rani, and my dearest siblings whose unwavering love, prayers, and encouragement have been my greatest source of inspiration and strength throughout this journey. Their endless support and sacrifices have been invaluable in helping me reach this milestone.

I would also like to extend my heartfelt gratitude to my fellow companions on this journey for their support and encouragement. Thank you for your companionship, motivation, and for standing by me through the challenges and triumphs of this PhD journey. Your kindness and support have made this experience even more meaningful.

Lastly, I extend my appreciation to myself for the perseverance, resilience, and dedication I have shown throughout this challenging yet rewarding path. Completing this PhD has been a true test of patience and determination, and I am proud of how far I have come.

Thank you for your support, encouragement, ideas and comments for the success of this project. This victory is dedicated to all of you. Alhamdulillah

TABLE OF CONTENTS

CHAPTER 1
INTRODUCTION

1.1 Research Background

According to the American Cancer Society (ACS) [1], an estimated 238,340 new lung cancer cases will be diagnosed in 2023, with approximately 127,070 deaths, accounting for 20% of all cancer-related fatalities. Lung cancer occurs when mutated cells multiply uncontrollably, forming tumours that invade and destroy surrounding tissues [2]. The two main types of lung cancer are small-cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC), with NSCLC being the most common and a leading cause of cancer-related deaths in both men and women in the United States. The incidence of lung cancer is rare in individuals under 40 but increases significantly in those over 70 [3]. Smoking remains the leading cause, responsible for nearly 80% of cases [1].

In Malaysia, the age-standardised incidence rate of lung cancer rises sharply after 45, peaking in the 60–74 age group, making it the third most common cancer in the country. A study from a Malaysian hospital found that all lung cancer patients presented with stage III or IV disease, with a median survival of just 18 weeks post-diagnosis. Advanced-stage diagnoses were reported in 93.5% of male and 92.3% of female patients. In 2020, Malaysia recorded 48,639 new cancer cases, and cancer incidence is expected to double by 2040. The 5-year observed survival rate for lung cancer is only 9.0% (95% confident interval (CI): 8.4-9.7), while the 5-year relative survival rate is 11.0% (95% CI: 10.3–11.9)[4], [5], [6].

Among imaging modalities, computed tomography (CT) scans are preferred for lung cancer screening due to their high sensitivity and ability to detect nodules as small as 4 mm [7] and [8]. The technology has significantly advanced since its introduction in the 1970s, with modern systems offering unmatched image quality [9]. Advanced Medical and Dental Institute (AMDI), Universiti Sains Malaysia (USM) has been chosen by Nucletron as a comprehensive cancer service provider and a center for therapeutic/diagnostic nuclear medicine and this highlight their cancer treatment capabilities in delivering advanced cancer services to the community [10]. At AMDI, USM, the Siemens SOMATOM Definition AS is utilised for lung CT imaging. This