# PREDICTION FOR HIGH-RISK SYMPTOMS OF LUNG CANCER IN MALAYSIA USING FUZZY LINEAR REGRESSION MODEL

ALIYA SYAFFA BINTI ZAKARIA

A thesis submitted in
fulfillment of the requirement for the
award of the Degree of Master of Science in Technology Management

Faculty of Technology Management & Business

SEPTEMBER 2023

# DEDICATION

*This dissertation would not be possible without every single encouragement I have received; therefore I dedicate this to them. A special feeling gratitude to my loving parents, Zakaria bin Abdul Rashid and Mazlina binti Basuri, that never stops making me believe myself, especially at times I feel the opposite and thank you for your patience for always striving to make me a better person.*

*My sister, Addina Iftitah, and my brother, Alfian Danial, that have never left my side and have a very special place in my heart.*

*I also dedicate to my big family for all your kindness throughout my degree process. I will always appreciate what you have done to give me extra support to keep moving in this research project.*

*Thank you to my best friend Muhammad Shazmin, that have supported me throughout the research process.*

# ACKNOWLEDGEMENTS

First and foremost, I would like to endeavor my gratitude to almighty God, Allah, for giving me good health, peace of mind, and strength to finish this research paper.

Furthermore, I would like to express sincere gratitude to my supervisor, Dr Muhammad Ammar bin Shafi, and my co-supervisor, Prof. Ts. Dr. Alina binti Shamsuddin, for the continuous support given to me for my research project. Thank you for your motivation, cooperation, and immense knowledge delivered to me. Your guidance helped me a lot in writing the research at all times.

Besides my supervisor, I also want to thank to the UiTM partners, Dr. Mohd Arif bin Mohd Zim, Nurnadiah Nordin, and Siti Mukhaiyarah Mahtar, for your help in giving the trust in me to generate the data obtained to complete my thesis.

# ABSTRACT

Lung cancer has been recorded as the most common cancer globally, contributing 12.2% of all new cases diagnosed in 2020, with the greatest mortality rate due to its late diagnosis and poor symptom detection. Nowadays, Malaysia has reached 4,319 lung cancer deaths, accounting for 2.57 per cent of all deaths in 2020. Late diagnosis is the norm for lung cancer, which makes survival challenging and the likelihood of recovery low. Nevertheless, in Malaysia, most cases are discovered late, when the tumors have grown too far, or the disease has spread to other body parts that cannot be removed through surgery. This situation frequently occurs due to the lack of public knowledge among Malaysians regarding cancer-related signs and symptoms. Therefore, Malaysians must be aware of the high-risk symptoms of lung cancer to increase the survival rate and decrease the mortality rate. This study aims to compare multiple linear regression and fuzzy linear regression model using a triangular fuzzy number proposed by Tanaka. The *H*-value from 0.0 to 1.0 is adjusted to find the optimal value of an objective function to predict high-risk lung cancer symptoms in Malaysia. The secondary data is analyzed using the fuzzy linear regression model, which can reduce the interference of irrelevant information and improve the precision of the results. This research data was collected from patients with lung cancer at Al-Sultan Abdullah Hospital (UiTM Hospital), Selangor. The data of 124 lung cancer patients were analyzed using Microsoft Excel and MATLAB. The study implemented measurement error of cross-validation technique, which is mean square error (MSE) and root mean square error (RMSE), to enhance data accuracy. The results show that haemoptysis and chest pain has been proven to be the highest risk, among other symptoms acquired from the data analysis. It has been determined that *H*-value of 0.0 has the smallest measurement error, with MSE of 1.455 and RMSE of 1.206 as the multiple linear regression method has the MSE value of 306.257 while the RMSE has the value of 17.500.

# ABSTRAK

Kanser paru-paru telah direkodkan sebagai kanser yang paling biasa di seluruh dunia, menyumbang 12.2% daripada semua kes baharu yang didiagnosis pada 2020, dengan kadar kematian yang paling tinggi disebabkan diagnosis lewat dan pengesanan simptom yang lemah. Hari ini, Malaysia telah mencapai 4,319 kematian akibat kanser paru-paru, menyumbang 2.57 peratus daripada semua kematian pada 2020. Diagnosis lewat adalah norma bagi kanser paru-paru, yang menjadikan kelangsungan hidup mencabar dan kemungkinan pemulihan rendah. Namun begitu, di Malaysia, kebanyakan kes ditemui lewat, apabila tumor telah membesar terlalu jauh, atau penyakit itu telah merebak ke bahagian badan lain yang tidak boleh dibuang melalui pembedahan. Keadaan ini kerap berlaku kerana kurangnya pengetahuan masyarakat Malaysia mengenai tanda dan gejala berkaitan kanser. Oleh itu, rakyat Malaysia mesti sedar tentang simptom berisiko tinggi kanser paru-paru untuk meningkatkan kadar kelangsungan hidup dan mengurangkan kadar kematian. Kajian ini bertujuan untuk membandingkan model regresi linear berganda dan regresi linear kabur menggunakan nombor kabur segi tiga yang dicadangkan oleh Tanaka. Nilai $H$ dari 0.0 hingga 1.0 dilaraskan untuk mencari nilai optimum bagi fungsi objektif untuk meramalkan gejala kanser paru-paru berisiko tinggi di Malaysia. Data sekunder dianalisis menggunakan model regresi linear kabur, yang boleh mengurangkan gangguan maklumat yang tidak berkaitan dan meningkatkan ketepatan keputusan. Data kajian ini dikumpul daripada pesakit kanser paru-paru di Hospital Al-Sultan Abdullah (Hospital UiTM), Selangor. Data daripada 124 pesakit kanser paru-paru dianalisis menggunakan Microsoft Excel dan MATLAB. Kajian ini melaksanakan ralat pengukuran teknik pengesahan silang, iaitu ralat min kuasa dua (MSE) dan ralat min kuasa dua punca (RMSE), untuk meningkatkan ketepatan data. Keputusan menunjukkan bahawa hemoptisis dan sakit dada telah terbukti sebagai risiko tertinggi, antara gejala lain yang diperoleh daripada analisis data. Telah ditentukan bahawa nilai $H$ 0.0 mempunyai ralat pengukuran terkecil, dengan MSE 1.455 dan RMSE 1.206 kerana kaedah regresi linear berganda mempunyai nilai MSE 306.257 manakala RMSE mempunyai nilai 17.500.

# CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# LIST OF SYMBOLS AND ABBREVIATION

| | | |
|---|---|---|
| AFDL | - | adaptive fuzzy dictionary learning |
| ANOVA | - | analysis of variance |
| BTS | - | baseline tumor size |
| CT | - | computed tomography |
| FBS | - | fibro bronchoscopy and biopsy |
| FLR | - | fuzzy linear regression |
| FLSA | - | fuzzy latent semantic analysis |
| GLOBOCAN | - | Global Cancer Observatory |
| ICI | - | immune checkpoint inhibitors |
| LDA | - | latent Dirichlet allocation |
| LDCT | - | low-dose computed tomography |
| MAE | - | mean absolute error |
| MAPE | - | mean absolute percentage error |
| Matlab | - | matrix laboratory |
| MLR | - | multiple linear regression |
| MSE | - | mean square error |
| MSR | - | mean square regression |
| NLST | - | National Lung Screening Trial |
| NSCLC | - | non-small cell lung cancer |
| PP-plot | - | plot of probability |
| QQ-plot | - | plot of quartile |
| RBFNN | - | radial basis function neural network |
| RMSE | - | root mean square error |
| SCLC | - | small cell lung cancer |
| SPSS | - | Statistical Package for Social Sciences |
| SSE | - | sums of error |
| SSR | - | sums of regression |
| SST | - | sums of total |
| SVD | - | singular value decomposition |
| TAA | - | tumor associated antigen |
| UiTM | - | Universiti Teknologi MARA |
| VIF | - | variation inflation factor |
| WHO | - | World Health Organization |

| | | |
|---|---|---|
| $\alpha$ | - | Center of fuzzy |
| $A_g$ | - | fuzzy parameter |
| $\beta_i$ | - | Coefficient in multiple linear regression |
| $\varsigma$ | - | width |
| $\varepsilon$ | - | Vector of independent normal random variables |
| $\Im$ | - | Infinite dimensional feature space |
| J | - | An nxn matrix |
| N | - | Number of observations 30 and above |
| $n_{th}$ | - | sample of *n* |
| $x_e$ | - | variable of fuzzy parameter |
| $\mu_A$ (a) | - | Membership function of element in set A |
| $x_n$ | - | input or independent variables for the sample |
| Y | - | Dependent variables/ observations |
| $Y^*_e$ | - | equation of the fuzzy parameter |
| $y_i$ | - | Observation of data |
| $\sum$ | - | Covariance matrix |
| $y_n$ | - | output or observation of data |

# LIST OF APPENDICES

# CHAPTER 1

# INTRODUCTION

## 1.1 Introduction

This chapter provides an overview of the research background of lung cancer, the introduction of fuzzy linear regression, problem statement, research questions and objectives, the scope and significance of the study. Finally, at the end of the chapter, there is a brief overview of the research organization.

## 1.2 Research Background of Lung Cancer

Cancer is a disease in which the body's cells proliferate uncontrollably. Lung cancer occurs when cancer begins in the lungs and can spread to lymph nodes or other organs such as the brain. Cancer from other organs can spread to the lungs as well. In 2020, the American Cancer Society reported that lung cancer (both small and non-small cell) is the second most prevalent cancer in both men and women (excluding skin cancer). The incidence of this type of cancer is on the rise in several countries, especially in Asian countries, where the rate went up from 56% in 2012 to 58% in 2018 (Pakzad *et al.*, 2015).

Lung cancer is the most common cause of cancer-related death in Malaysia. The data of 2020 from the World Health Organization (WHO) showed that lung cancer is the leading cause of cancer mortalities with 1.80 million deaths, followed by colon and rectum cancer with 935 000 deaths, and liver cancer with 830 000 deaths. While according to the most recent WHO data published in 2020, Malaysia has reached a number of 4,319 lung cancer deaths, accounting for 2.57 percent of all deaths. Malaysia is the 77$^{th}$ country in the world with an age-adjusted mortality rate of 15.25 per 100,000 population. The reported 5-year survival rate is only 9.0% (95% confidence interval: 8.4–9.7), however, the relative 5-year survival rate is 11.0% (95% confidence interval: 10.3–11.1). As depicted in Chart A, the survival rate of lung cancer patients in Malaysia at 1 and 5 years is one of the lowest compared to other cancer types. Chart B displays the 1-year and 5-year survival rates by stage (National Cancer Registry,2018).

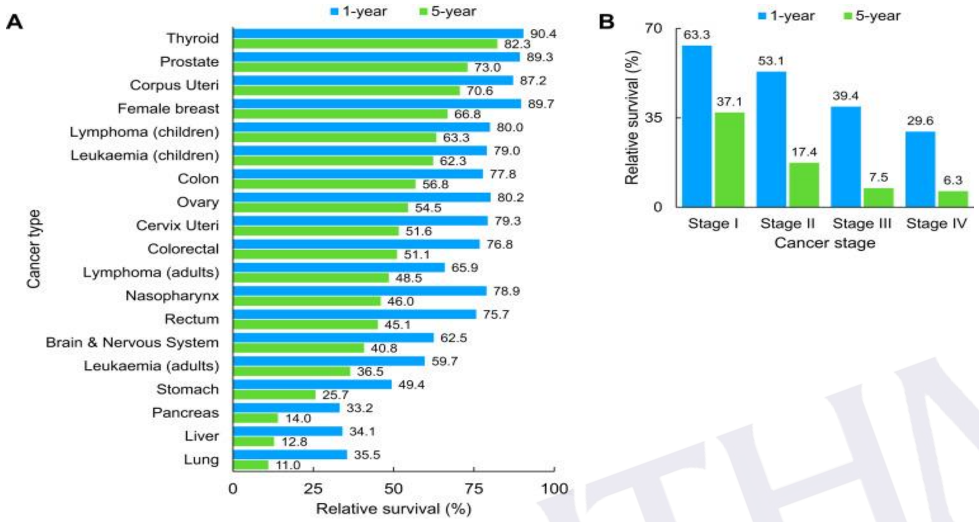


Figure 1.1: Relative survival of cancer patients in Malaysia

Lung cancer can be diagnosed through symptoms or related signs such as coughing up blood, weakness, weight loss, fever or clubbing of the fingernails, my asthenia syndrome (muscle weakness), hypercalcemia, and metastases. At the same time, the most frequent clinical signs are coughing (including coughing of blood), weight loss, shortness of breath and chest pain. Early detection of lung cancer may be an effective strategy for improving patient care, resulting in a lowered mortality rate. Additionally, National Lung Screening Trial (NLST) reported that early screening reduced the mortality rate of lung cancer by 20% (Midthun, 2016). The current approaches to lung cancer screening include X-rays and computed tomography (CT). However, the reliability of these approaches is dubious, as the false positive rate in these trials exceeded about 15% (Knight *et al.*, 2017). Thus, several alternative methods, such as metabolomic, transcriptomic, genomic, and proteomic for identifying cancerous biomarkers for the early detection of lung cancer have been studied recently (Jalal *et al.*, 2021).

According to American Lung Association (2021), lung cancer consists of two kinds of cells which are small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC). SCLC is delineated using two stages, namely limited and extensive. NSCLC stages range from one to four, commonly denoted in Roman numerals (0 through IV). Stage 0 indicates cancer only in the top lining of the lung and has not spread. Cancer has not spread to the lymph nodes or other parts of the body in Stage I, while the tumors may be larger than those in Stage I and/or have started to spread to nearby lymph nodes in Stage II. Stage III can be determined when cancer has spread to the mediastinum lymph nodes (the chest area between the lungs). In Stage IV, cancer has metastasized or spread in the lining of the lung or other body areas.

In Malaysia, about 1 in 60 males and 1 in 138 females develop lung cancer, while the mean age for lung cancer is 70 years and above (range 15 to 90 years). However, most lung cancer cases were detected at a very late stage, Stage III and IV, which is above 90% for both sexes. Patients with lung cancer have two various ways of presentation for diagnosis; symptomatic and incidental. Most patients were diagnosed incidentally on chest X-rays and CT scans. While amongst the patients who had symptoms, the most frequently reported complaints that resulted in an imaging referral were the development of new cough or the incline of a previously manifested clinical picture suggestive of pneumonia and haemoptysis (Quadrelli *et al.*, 2015).

Figure 1.2 shows the tumor sites and Figure 1.3 displays the stages of lung cancer.

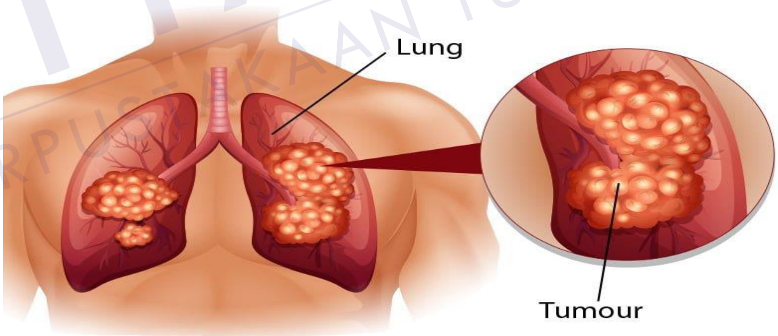


Figure 1.2: Lung cancer sites

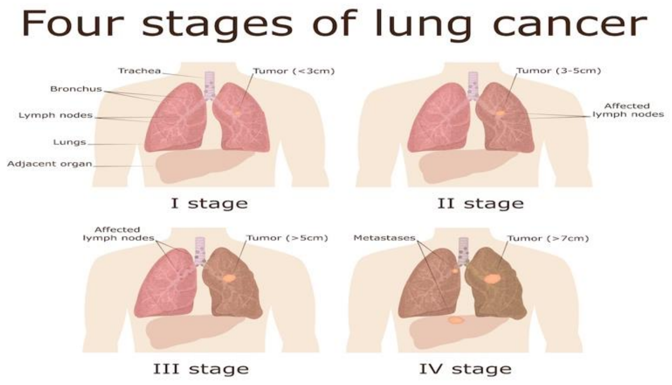


Figure 1. 3: Stages of lung cancer

## 1.3 Introduction of Fuzzy Linear Regression Analysis

Fuzzy linear regression analysis is an essential alternative to frequently used statistics-based regression methods. An ample range of fuzzy linear models can be applied in fuzzy linear regression analysis to approximate a linear dependence based on a set of observations. Fuzzy regression consists of two types. Tanaka *et al.* (1982) developed 'possibilistic' fuzzy regression, which is known as a linear programming method that seeks to reduce the system's fuzziness. The second method is a fuzzy least-squares method, which seeks to minimize the distance between two fuzzy numbers. The approaches are designed to fit fuzzy data to meet a specific applicable requirement. (Khan & Valeo, 2015).

Linear regression analysis with a fuzzy model was introduced by Tanaka *et al.* (1982). The structure is expressed by fuzzy sets as a fuzzy linear function. Zadeh's extension principle proposed fuzzy linear functions. They used input and output data on property prices and a fuzzy linear regression model to analyze the data. Fuzzy linear functions were found to be a validated strategy for dealing with ambiguous occurrences when applied in the linear regression model.

A fuzzy regression model is used to determine the functional relationship between the dependent and independent variables in a fuzzy environment. Numerous fuzzy regression models have been proposed in the literature and several approaches for estimating the models' fuzzy parameters. The possibilistic approach and the fuzzy least squares model are the two most common methods in analyzing fuzzy regression models (Denoda *et al.*, 2014).

Poleshchuk (2018) developed an output variable in a fuzzy linear regression model corresponding to confident intervals with a specified level of plausibility. The fuzzy regression analysis methods extend the classical regression analysis methods by solving various problems with fuzzy or incomplete initial data without resorting to the methods of probabilities. The proposed method opens up new potential for predicting fuzzy output variables.

Gkountakou & Papadopoulos (2020) applied fuzzy linear regression to construct more effective fuzzy models for estimating cement's 28-day compressive strength. The fuzzy linear regression approach is a powerful tool for defining the degree of fuzziness and calculating the effect of independent inputs on the dependent variable. It also establishes a standard equation for estimating output values through symmetric triangular fuzzy numbers and identifies the most critical component in enhancing compressive strength. It is proven that fuzzy linear regression is an appropriate strategy for engineering mathematical models through fuzzy logic.

Fuzzy linear regression is an intriguing and potentially practical strategy for overcoming this gap. Additionally, this fuzzy methodology is capable of effectively dealing with the issue of multicollinearity. Pandit *et al.*, (2021) indicated that fuzzy methodology clearly outperforms conventional regression methodology when many interconnected factors are required to forecast an outcome variable. It is also clearly demonstrated that fuzzy linear regression has a higher relative efficiency than simple and multiple linear regression methods.

Fuzzy linear regression is well-suited for vague data in modeling. Clustering is applied to group or cluster data based on similarities, where fuzzy C-means is the best method. Fuzzy C-means clustering can be classified as the best method since it can handle big datasets and allows an item to belong to more than one cluster. Ramly *et al.*, (2018) proved that fuzzy C-means and fuzzy linear regression models as the best techniques for predicting manufacturing income. This is because the improvisation model obtains the lowest mean square error (MSE), mean absolute error (MAE), and mean absolute percentage error (MAPE) than other models, such as multiple linear regression.

## 1.4 Problem Statement

This research is carried out to predict the high-risk signs and symptoms of people who have early detection of lung cancer. Numerous research has been conducted on lung cancer (Yu, 2019). However, studies on the early detection of high-risk symptoms of lung cancer (stage I and II) remain inconclusive and unspecific. Currently, the Malaysian Health Technology Assessment Section reported that low dose computed tomography (LDCT) was used for lung cancer screening and improved lung cancer detection. However, it does not apply to high-risk lung cancer patients. Identifying the high-risk individual with screening outcomes is crucial since this increases the likelihood of detecting malignancy early and reduces lung cancer mortality (Wille *et al.*, 2015).

According to Sachithanandan & Badmanaban (2012), early-stage disease (I, II, and selected IIIa) is susceptible to curative surgery, providing the best chance of cure and disease-free survival in the long term. Nevertheless, in Malaysia, most cases are discovered late, when the tumors have grown too far or the disease has spread to other body parts that cannot be removed through surgery. This situation frequently occurs due to the significant gaps in the public's knowledge of Malaysian people regarding cancer-related signs, symptoms, and factors (Schliemann, 2020). The majority of patients (roughly 75 per cent) have an advanced illness at the time of diagnosis (stage III/IV). Despite substantial breakthroughs in late-stage lung cancer oncological care in latest years, survival remains low (Knight *et al.*, 2017).

Regression analysis is a statistical methodology applied to determine the relationship between two variables with a cause-and-effect relationship. Regression analysis can be a powerful tool for understanding (including predicting and explaining) the causal influence on a population outcome (Jihye, 2015). However, regression models are particularly sensitive to outliers. A data point that deviates greatly from the majority of other observations is called an outlier. Variability in measurement may cause an experimental error, while an outlier in regression analysis might create a significant problem. Regression analysis models also over-simplify the actual world data and problems as the data is rarely linearly separable.

Moreover, Al-Sabri (2020) found that the fuzzy linear regression method is easier and clearer to calculate than the classical regression, and it does not differ significantly from the classical regression. These findings also lend support to the concept of fuzzy linear regression prediction, particularly in relation to fuzzy data. The high-risk symptoms of lung cancer can be determined more accurately by applying the fuzzy linear regression method, as it can predict the uncertainty data clearer compared to regression analysis.

## 1.5 Research Questions

i. How to analyze the multiple linear regression and fuzzy linear regression model in predicting high-risk symptoms of lung cancer in Malaysia?
ii. How does the performance of the multiple linear regression and fuzzy linear regression model be measured to determine the optimal model for predicting high-risk symptoms of lung cancer?
iii. What is the highest-risk symptom that most significantly impacts lung cancer symptoms in terms of early detection?

## 1.6 Research Objectives

The objectives of this research are;

i. To analyze fuzzy linear regression model by adjusting *H*-value in predicting high-risk symptoms of lung cancer in Malaysia.
ii. To measure the performance of fuzzy linear regression and multiple linear regression models using statistical measurement errors such as mean square error (MSE) and root mean square error (RMSE).
iii. To determine which symptoms have the most significant impact on the symptoms of lung cancer using fuzzy linear regression method.

## 1.7 The Scope of Study

The scope of the study will be elucidated in two parts; the data scope and the model scope.

### 1.7.1 Data Scope

This research will cover the respondents, types of data obtained and the fuzzy linear regression model. The study population will consist of patients diagnosed with lung cancer at all stages who attend a respiratory appointment at Hospital Al-Sultan Abdullah (UiTM Hospital) in Malaysia. This research mainly will focus on the multiple linear regression and fuzzy linear regression model to prove that fuzzy linear regression has the least measurement error in predicting high-risk symptoms of lung cancer. Secondary data will be used in this research, where the real data on lung cancer were acquired from Hospital Al-Sultan Abdullah (UiTM Hospital), Selangor. The respondents with lung cancer involved 124 patients, and the data were collected and recorded by doctors and nurses using cluster sampling. As continuous data, the dependent variable and independent variable are included. The dependent variable is the tumor size, while the independent variables are the symptoms of lung cancer, namely cough, haemoptysis, weight loss, loss of appetite, chest pain, smoking habit and comorbidity as continuous and categorical data included. SPSS, MATLAB and Microsoft Excel are potential soft computing software that will be applied to provide precise results.

### 1.7.2 Model Scope

Tanaka et al. (1982) developed fuzzy linear regression, which utilizes fuzzy parameters to model the vague and inaccurate relationship between dependent and independent variables. The main benefit of the Tanaka model is its simplicity in computation. This research proposed the multiple linear regression and fuzzy linear regression with *H*-values of 0.0 until 1.0 to compare both models in terms of error values. By applying this model in the medical field, it is anticipated to predict high-risk symptoms of lung cancer at any stage more accurately. This application model consists of several procedures, which will be discussed in further detail in Chapter 3 of the study.

### 1.7.3 Variables of the Study

The dependent variable is tumor size. The examined tumor ranges in size from 30 mm to 100 mm. The medical doctors asked the feedback about the patient's health and received responses promptly during the appointment session. Tumor size is chosen as the dependent variable since it can determine a patient's lung cancer stage from stage I until IV. The symptoms of lung cancer are likely to appear more when the stage is higher. While symptoms listed below have been chosen based on the cluster sampling method by doctors and nurses in UiTM Hospital which clustered the information from all types of cancer, such as colorectal, lung, and breast cancer. The questions addressed eleven factors except for cancer stages. Not only have the symptoms been faced by most lung cancer patients in the data collection, but they also mainly were stated by past researchers that studied the symptoms of lung cancer. The study of past researchers' findings on the chosen lung cancer symptoms will be elaborated more in Chapter 2.

Table 1.1: Description of data

| No | Variable name | Variable Type | Note |
|---|---|---|---|
| 1 | Tumor size | Quantitative discrete with Minimum size = 30mm and Maximum size = 100mm | Size of tumor by patient |
| 2 | Gender | Qualitative binary with Female and Male | Gender of patient |
| 3 | Age | Quantitative discrete with Minimum age = 47 and Maximum age = 95 | Age of patient |
| 4 | Ethnic | Qualitative categories with Malay, Chinese, Indian and Non-citizen | Ethnic of patient |
| 5 | Cough | Qualitative binary with Yes and No | Symptom faced by patient |
| 6 | Haemoptysis | Qualitative binary with Yes and No | Patient who suffers from coughing up blood from the lungs or bronchial tubes |

**REFERENCES**

Abadi, A. M., Wutsqa, D. U., & Ningsih, N. (2021). Construction of fuzzy radial basis function neural network model for diagnosing prostate cancer. *Telkomnika (Telecommunication Computing Electronics and Control)*, *19*(4), 1273–1283.

Abdel-Rahman, O. (2020). *Incidence and Mortality of Lung Cancer Among Never 2Smokers in Relationship to Secondhand Smoking: Findings From the PLCO Trial. Clinical Lung Cancer*, *21*(5), 415-420.

Agresti, A. (1996). An Introduction to Categorical Data Analysis New York : John Wiley & Sons, Inc, 1-14. Abdel-Rahman, O. (2020). Incidence and Mortality of Lung Cancer Among Never Smokers in Relationship to Secondhand Smoking: Findings From the PLCO Trial. *Clinical Lung Cancer*, *21*(5), pp. 415-420.

Arooj, P., Bredin, E., Henry, M. T., Khan, K. A., Plant, B. J., Murphy, D. M., & Kennedy, M. P. (2018). Bronchoscopy in the investigation of outpatients with hemoptysis at a lung cancer clinic. *Respiratory Medicine*, *139*, pp. 1–5.

Athey, V. L., Walters, S. J., & Rogers, T. K. (2018). Symptoms at lung cancer diagnosis are associated with major differences in prognosis. *Thorax*, *73*(12), pp. 1177–1181.

Bade, B. C., & Dela Cruz, C. S. (2020). Lung Cancer 2020: Epidemiology, Etiology, and Prevention. *Clinics in Chest Medicine*, *41*(1), pp. 1–24.

Bajpai, A., & Kushwah, V. S. (2019). Importance of fuzzy logic and application areas in engineering research. *International Journal of Recent Technology and Engineering*, *7*(6), pp. 1467–1471.

Bankar, A., Padamwar, K., & Jahagirdar, A. (2020). Symptom analysis using a machine learning approach for early stage lung cancer. *Proceedings of the 3rd International Conference on Intelligent Sustainable Systems, ICISS 2020*, pp. 246–250.

Bonney, A., Malouf, R., Marchal, C., Manners, D., Fong, K. M., Marshall, H. M., Irving, L. B., & Manser, R. (2022). Impact of low-dose computed tomography (LDCT) screening on lung cancer-related mortality. *Cochrane Database of Systematic Reviews*, *2022*(8).

Chowienczyk, S., Price, S., & Hamilton, W. (2020). *Changes in the presenting symptoms of lung cancer from 2000 – 2017 : a serial cross-sectional study of observational records in UK primary care*. *March*, 193–199.

Fayek, A. R., & Lourenzutti, R. (2018). Introduction to Fuzzy Logic in Construction

Engineering and Management. In *Fuzzy Hybrid Computing in Construction Engineering and Management*.

Feng, Y., Dai, W., Wang, Y., Liao, J., Wei, X., Xie, S., Xu, W., Li, Q., Liu, F., & Shi, Q. (2021). Comparison of chief complaints and patient-reported symptoms of treatment-naive lung cancer patients before surgery. *Patient Preference and Adherence*, *15*, pp. 1101–1106.

Franceschini, J. P., Jamnik, S., & Santoro, I. L. (2020). Role that anorexia and weight loss play in patients with stage iv lung cancer. *Jornal Brasileiro de Pneumologia*, *46*(4), pp. 1–6.

Galvez, M., Rossana, N., Joseph, R., Katia, A. P., Raul, R., & Luis, M. (2019). Lung Cancer in the Young. *Lung*.

Gkountakou, F., & Papadopoulos, B. (2020). The use of fuzzy linear regression and ANFIS methods to predict the compressive strength of cement. *Symmetry*, *12*(8).

González Maldonado, S., Motsch, E., Trotter, A., Kauczor, H. U., Heussel, C. P., Hermann, S., Zeissig, S. R., Delorme, S., & Kaaks, R. (2021). Overdiagnosis in lung cancer screening: Estimates from the German Lung Cancer Screening Intervention Trial. *International Journal of Cancer*, *148*(5), pp. 1097–1105.

Haastrup, P. F., Jarbol, D. E., Balasubramaniam, K., Saetre, L. M. S., Sondergaard, J., & Rasmussen, S. (n.d.). Predictive values of lung cancer alarm symptoms in the general population : a nationwide cohort study. *Npj Primary Care Respiratory Medicine*, *400*(Table 1), pp. 1–7.

Hamada, T., Komatsu, H., Rosenzweig, M. Q., Chohnabayashi, N., Nishimura, N., Oizumi, S., & Ren, D. (2016). Impact of symptom clusters on quality of life outcomes in patients from japan with advanced nonsmall cell lung cancers. *Asia-Pacific Journal of Oncology Nursing*, *3*(4), pp. 370–381.

Harle, A., Blackhall, F. H., Molassiotis, A., Yorke, J., Dockry, R., Holt, K., Yuill, D., Baker, K., & Smith, J. A. (2018). Cough in Patients With Lung Cancer A Longitudinal Observational Study of Characterization and Clinical Associations. *CHEST*, *October*, pp. 1–11.

Huang, K. L., Wang, S. Y., Lu, W. C., Chang, Y. H., Su, J., & Lu, Y. T. (2019). Effects of low-dose computed tomography on lung cancer screening: A systematic review, meta-analysis, and trial sequential analysis. *BMC Pulmonary Medicine*, *19*(1), pp. 1–11.

Jalal, A. H., Sikder, A. K., Alam, F., Samin, S., Rahman, S. S., Khan, M. M., & Siddiquee, M. R. (2021). Early diagnosis with alternative approaches: Innovation in lung cancer care. *Shanghai Chest*, *5*, 7-7.

Kaaks, R., Christodoulou, E., Motsch, E., Katzke, V., Wielpütz, M. O., Kauczor, H. U., Heussel, C. P., Eichinger, M., & Delorme, S. (2022). Lung function impairment in the German Lung Cancer Screening Intervention Study (LUSI): prevalence, symptoms, and associations with lung cancer risk, tumor histology and all-cause mortality. *Translational Lung Cancer Research*, *11*(9), pp. 1896–1911.

Kraft, D., Bordogna, G., & Pasi, G. (2021). Fuzzy Set Theory. *Encyclopedia of Computer Science and Technology*, pp. 441–458.

Kumari, Khushbu & Yadav, Suniti. (2018). Linear regression analysis study. Journal of the Practice of Cardiovascular Sciences. 4. 33.

Lasake, I. B., Idayu, R., Mat, B., Binti, N., & Marzuki, M. (2020). *Recurrent Haemoptysis in Non-Small Cell Lung Cancer Patient* . *2*(1), pp. 18–19.

Lebrett, M. B., Balata, H., Evison, M., Colligan, D., Duerden, R., Elton, P., Greaves, M., Howells, J., Irion, K., Karunaratne, D., Lyons, J., Mellor, S., Myerscough, A., Newton, T., Sharman, A., Smith, E., Taylor, B., Taylor, S., Walsham, A., … Crosbie, P. A. J. (2020). *Analysis of lung cancer risk model ( PLCO M2012 and LLP v2 ) performance in a community- - based lung cancer screening programme*. pp. 1–8.

Levitsky, A., Pernemalm, M., Bernhardson, B. M., Forshed, J., Kölbeck, K., Olin, M., Henriksson, R., Lehtiö, J., Tishelman, C., & Eriksson, L. E. (2019). Early symptoms and sensations as predictors of lung cancer: a machine learning multivariate model. *Scientific Reports*, *9*(1), pp. 1–11.

Loh, J. F., & Tan, S. L. (2018). Lung cancer knowledge and screening in the context of the Malaysian population. *Journal of Pharmacy Practice and Research*, *48*(1), pp. 56–64.

M.A., S. D., & N.Ch.S.N, I. (2010). *Effective analysis and diagnosis of lung cancer*. *2*(6), pp. 2102–2108.

Mansour, A. I., Abu-Jamie, T. N., Al-Masawabe, M. M., & ... (2021). *An Expert System for Diagnosing Cough Using SL5 Object*. *5*(5), pp. 79–90.

Mccutchan, G., Smits, S., Ironmonger, L., Slyne, C., Boughey, A., Moffat, J., Thomas, R., Huws, D. W., & Brain, K. (2019). Evaluation of a national lung cancer symptom awareness campaign in Wales. *British Journal of Cancer*, *December*.

Miranda-Filho, A., Piñeros, M., & Bray, F. (2019). The descriptive epidemiology of lung cancer and tobacco control: A global overview 2018. *Salud Publica de Mexico*, *61*(3), pp. 219–229.

Mondoni, M., Carlucci, P., Cipolla, G., Pagani, M., Tursi, F., Fois, A., Pirina, P., Canu, S., Gasparini, S., Bonifazi, M., Marani, S., Comel, A., Saderi, L., Pascalis, S. De, Alfano, F., Centanni, S., & Sotgiu, G. (2021). Long - term prognostic outcomes in patients with haemoptysis. *Respiratory Research*, pp. 1–7.

Morishima, T., Matsumoto, Y., Koeda, N., Shimada, H., Maruhama, T., Matsuki, D., Nakata, K., Ito, Y., Tabuchi, T., & Miyashiro, I. (2019). *Journal of Epidemiology*. *29*(3), pp. 110–115.

Munawar, Z., Ahmad, F., Awadh Alanazi, S., Nisar, K. S., Khalid, M., Anwar, M., & Murtaza, K. (2022). Predicting the prevalence of lung cancer using feature transformation techniques. *Egyptian Informatics Journal*, *23*(4), pp. 109–120.

Nasrullah, N., Sang, J., Alam, mohammad s., Mateen, M., Cai, B., & Hu, H. (2019). Automated lung nodule detection and classification using deep learning combined with multiple strategies. *Microscopy Research and Technique*, *82*(9), pp. 1601–1609.

Nasser, I. M., & Abu-Naser, S. S. (2019). *Lung cancer Detection using Artificial Neural Network*. *3*(3), pp. 17–23.

Nie, L., Dai, K., Wu, J., Zhou, X., Hu, J., Zhang, C., Zhan, Y., Song, Y., Fan, W., Hu, Z., Yang, H., Yang, Q., Wu, D., Li, F., Li, D., & Nie, R. (2021). Clinical characteristics and risk factors for in-hospital mortality of lung cancer patients with COVID-19: A multicenter, retrospective, cohort study. *Thoracic Cancer*, *12*(1), pp. 57–65.

Petersen, C. L., & Weinreich, U. M. (2019). Five-year follow-up of hemoptysis with no malignancy suspected on chest computed tomography: recurrence, lung cancer and mortality. *European Clinical Respiratory Journal*, *6*(1).

Polanski, J., Jankowska-Polanska, B., Rosinczuk, J., Chabowski, M., & Szymanska-Chabowska, A. (2016). Quality of life of patients with lung cancer. *OncoTargets and Therapy*, *9*, pp. 1023–1028.

Poleshchuk, O. M. (2018). Confidence intervals for output variable in fuzzy linear regression model. *IOP Conference Series: Materials Science and Engineering*, *468*(1), pp. 1–7.

Quigley, N., Gagnon, S., & Fortin, M. (2020). Aetiology, diagnosis and treatment of moderate-to-severe haemoptysis in a North American academic centre. *ERJ Open Research*, *6*(4), pp. 204–2020.

Ramly, N., Rusiman, M. S., Him, N. C., Nor, M. E., Man, S., Basri, N. Z. A., & Mohamad, N. (2018). A new hybrid of Fuzzy C-Means Method and Fuzzy Linear Regression Model in Predicting Manufacturing Income. *International Journal of Engineering & Technology*, *7*(43), pp. 473.

Raoof, S. S., Jabbar, M. A., & Fathima, S. A. (2020). Lung Cancer Prediction using Machine Learning: A Comprehensive Approach. *2nd International Conference on Innovative Mechanisms for Industry Applications, ICIMIA 2020 - Conference Proceedings*, *Icimia*, pp. 108–115.

Refaeilzadeh, P., Tang, L., Liu, H., Angeles, L., & Scientist, C. D. (2020). Encyclopedia of Database Systems. *Encyclopedia of Database Systems*.

Rodriguez, A. M., Braverman, J., Aggarwal, D., Friend, J., & Duus, E. (2017). The experience of weight loss and its associated burden in patients with non-small cell lung cancer: results of an online survey. *JCSM Clinical Reports*, *2*(2), 1–12.

Ruano-Raviña, A., Provencio, M., Calvo De Juan, V., Carcereny, E., Moran, T., Rodriguez-Abreu, D., López-Castro, R., Cuadrado Albite, E., Guirado, M., Gómez González, L., Massutí, B., Ortega Granados, A. L., Blasco, A., Cobo, M., Garcia-Campelo, R., Bosch, J., Trigo, J., Juan, Ó., Aguado De La Rosa, C., … Cerezo, S. (2020). Lung cancer symptoms at diagnosis: Results of a nationwide registry study. *ESMO Open*, *5*(6), 1–7.

Rudin, C. M., Brambilla, E., Faivre-Finn, C., & Sage, J. (2021). Small-cell lung cancer. *Nature Reviews Disease Primers*, *7*(1).

Schmidheiny, K. (2021). The Multiple Linear Regression Model 1. *An Introduction to Statistical Methods and Data Analysis Sixth Edition*, 664–762.

Seong, G. M., Hyun, C. L., Lee, J., & Kim, C. (2020). Large cell carcinoma of the lung presenting as diffuse pulmonary infiltrates with haemoptysis. *Respirology Case Reports*, *8*(7), 1–4.

Shankar, A., Saini, D., Dubey, A., Roy, S., Bharati, S. J., Singh, N., Khanna, M., Prasad, C. P., Singh, M., Kumar, S., Sirohi, B., Seth, T., Rinki, M., Mohan, A., Guleria, R., & Rath, G. K. (2019). Feasibility of lung cancer screening in developing countries: Challenges, opportunities and way forward. *Translational Lung Cancer Research*, *8*(Suppl 1), S106–S121.

Subha, S. T. (2019). Cancer: Its alarming trends. *Malaysian Journal of Medicine and Health Sciences*, *15*(3), pp. 1–3.

Sung, M. R., & Leighl, N. B. (2019). Improving lung cancer diagnosis: The evolving role of patients and care providers. *Journal of Thoracic Disease*, *11*(Suppl 3), pp. 422–424.

Teh, H. S., & Woon, Y. L. (2021). Burden of cancers attributable to modifiable risk factors in Malaysia. *BMC Public Health*, Vol. *21*, No. 1, pp. 1–10.

Thabit, H., & Zainuddin, N. (2017). Knowledge and perception on lung cancer and its screening: a study among undergraduate students of the International Islamic University Malaysia, Kuantan Campus. *Journal of Biomedical and Clinical Sciences*, Vol. *2, No.* 2, pp. 61–66.

Tirzite, M., Bukovskis, M., Strazda, G., Jurka, N., & Taivans, I. (2017). Detection of lung cancer in exhaled breath with an electronic nose using support vector machine analysis. *Journal of Breath Research,* Vol. *11, No.* 3.

Wang, J., Ran, F., Li, X., Bruera, E., & Qian, Y. (2022). *Dapagliflozin Related Severe Weight Loss in a Patient with Lung Cancer*. Vol. *7*, pp. 7–9.

Watte, G., Nunes, C. H. de A., Sidney-Filho, L. A., Zanon, M., Altmayer, S. P. L., Pacini, G. S., Barros, M., Moreira, A. L. S., Alves, R. J. V., Zelmanowicz, A. de M., Matata, B. M., & Moreira, J. da S. (2018). Proportional weight loss in six months as a risk factor for mortality in stage IV non-small cell lung cancer. *Jornal Brasileiro de Pneumologia*, *44*(6), pp. 505–509.

Whisenant, M. S., Williams, L. A., Gonzalez, A. G., & Mendoza, T. (2022). *What Do Patients With Non – Small-Cell Lung Cancer Experience ? Content Domain for the MD Anderson Symptom Inventory for Lung Cancer What Do Patients With Non – Small-Cell Lung Cancer Experience ? Content Domain for the MD Anderson Symptom Inventory for L*. Vol. *16*, No. 10.

Wille, M., Dirksen, A., Ashraf, H., Saghir, Z., Bach, K., & Brodersen, J. et al. (2016). Results of the Randomized Danish Lung Cancer Screening Trial with Focus on High-Risk Profiling. *American Journal Of Respiratory And Critical Care Medicine*, *193*(5), 542-551.

Wu, X., Denise, B.-B., Zhan, F., & Zhang, J. (2022). Determining Association between Lung Cancer Mortality Worldwide and Risk Factors Using Fuzzy Inference Modeling and Random Forest Modeling. *International Journal of Environmental Research and Public Health*, Vol. *19*, No. 21, pp. 14161.

Xing, P. Y., Liu, S. M., Du, L. Bin, Zhang, K., Zhang, Y. Z., Qiao, Y. L., & Wang, D. Bin. (2019). *What are the clinical symptoms and physical signs for non - small cell lung cancer before diagnosis is made ? A nation - wide multicenter 10 - year retrospective study in China*. *April*, pp. 4055–4069.

Yang, C. H., Moi, S. H., Hou, M. F., Chuang, L. Y., & Lin, Y. Da. (2020). Applications of Deep Learning and Fuzzy Systems to Detect Cancer Mortality in Next-Generation Genomic Data. *IEEE Transactions on Fuzzy Systems*, *29*(12), 3833–3844.

Yılmaz, A., Arı, S., & Kocabıçak, Ü. (2016). Risk analysis of lung cancer and effects of stress level on cancer risk through neuro-fuzzy model. *Computer Methods and Programs in Biomedicine*, *137*, 35–46.

Zadeh, L. A. (1965). Fuzzy sets. *Inform Control*, (8), pp. 338-358.

Zeng, W., Feng, Q., & Li, J. (2017). Fuzzy least absolute linear regression. *Applied Soft Computing Journal*, *52*, 1009–1019.

# APPENDIX A

## List of Publications

The following papers have been accepted and published were completed during master candidature.

**Refereed Journals: International (Submitted paper)**

Aliya Syaffa Zakaria, Muhammad Ammar Shafi, Mohd Arif Mohd Zim, Nurnadiah Nordin and Siti Mukhaiyarah Mahtar (2022). Computation of Fuzzy Linear Regression Model Using Simulation Data. *International Conference on Global Optimization and Its Application 2022 (ICoGOIA 2022)(Scopus)*

Aliya Syaffa Zakaria, Muhammad Ammar Shafi, Mohd Arif Mohd Zim, Aisya Natasya Musa (2023). Comparison of prediction fuzzy modeling towards high-risk symptoms of lung cancer. *Journal Of Intelligent and Fuzzy Systems. (Scopus)*

**Refereed Journals: International (Published paper)**

Aliya Syaffa Zakaria, Muhammad Ammar Shafi, Mohd Arif Mohd Zim, Siti Noor Asyikin Mohd Razali (2023). The use of fuzzy linear regression modeling to predict high-risk symptoms of lung cancer in Malaysia. *International Journal of Advanced Computer Science and Applications, Vol. 14, No. 5 (IJACSA 2023)(Scopus)*

# VITA

The author was born in Kedah, Malaysia, on October 07, 1998. Her secondary school was Sekolah Menengah Sungai Layar in Sungai Petani, Kedah, Malaysia. 1998 marked his graduation from the Universiti Tun Hussein Onn Malaysia, Johor, Malaysia, with a B.Eng. (Hons) in Technology Management and Business. She was actively participated in Sekretariat Rakan Muda club when she pursued her Degree study. She was graduated as a First Class Degree and been received a Vice Chancellor award. She then enrolled at Universiti Tun Hussein Onn Malaysia as Graduate Research Assistant while pursuing her Master study. During her Master study, she had won three awards; Silver award in Research And Innovation Symposium And Exposition 2022 competition, Silver award in Virtual Innovation Competition 2022 competition and runner-up position in SCIEMATHIC 2022 and achieved IP protection for her research. She also had secured a Memorandum of Understanding (MoU) with Universiti Teknologi Mara (UiTM) for the research.