The Use of Fuzzy Linear Regression Modeling to Predict High-risk Symptoms of Lung Cancer in Malaysia

Aliya Syaffa Zakaria[1], Muhammad Ammar Shafi[2], Mohd Arif Mohd Zim[3], Siti Noor Asyikin Mohd Razali[4]
Department of Technology and Management-Faculty of Technology Management and Business, Universiti Tun Hussein Onn Malaysia, 86400 Batu Pahat, Johor, Malaysia[1, 2]
Consultant Pulmonologist & Internal Medicine, Damansara Specialist Hospital 2, Jalan Bukit Lanjan 3, Bukit Lanjan, 60000 Kuala Lumpur, Malaysia[3]
Department of Mathematics and Statistics-Faculty of Applied Science and Technology, Universiti Tun Hussein Onn Malaysia, Pagoh Education Hub, 84600 Pagoh, Johor, Malaysia[4]

***Abstract*—Lung cancer is the most prevalent cancer in the world, accounting for 12.2% of all newly diagnosed cases in 2020 and has the highest mortality rate due to its late diagnosis and poor symptom detection. Currently, there are 4,319 lung cancer deaths in Malaysia, representing 2.57 percent of all mortality in 2020. The late diagnosis of lung cancer is common, which makes survival more difficult. In Malaysia, however, most cases are detected when the tumors have become too large, or cancer has spread to other body areas that cannot be removed surgically. This is a frequent situation due to the lack of public awareness among Malaysians regarding cancer-related symptoms. Malaysians must be acknowledged the high-risk symptoms of lung cancer to enhance the survival rate and reduce the mortality rate. This study aims to use a fuzzy linear regression model with heights of triangular fuzzy by Tanaka (1982), *H*-value ranging from 0.0 to 1.0, to predict high-risk symptoms of lung cancer in Malaysia. The secondary data is analyzed using the fuzzy linear regression model by collecting data from patients with lung cancer at Al-Sultan Abdullah Hospital (UiTM Hospital), Selangor. The results found that haemoptysis and chest pain has been proven to be the highest risk, among other symptoms obtained from the data analysis. It has been discovered that the *H*-value of 0.0 has the least measurement error, with mean square error (MSE) and root mean square error (RMSE) values of 1.455 and 1.206, respectively.**

Keywords*—Lung cancer; high-risk symptom; fuzzy linear regression; H-value; mean square error

I. Introduction

Cancer is a disease caused by uncontrolled cell division. Lung cancer develops when cancer originates in the lungs and spreads to lymph nodes or other organs, such as the brain. Moreover, lung cancer may spread from other organs. Lung cancer includes four stages which in Stage I, cancer has not grown to lymph nodes or other parts of the body, whereas in Stage II, the tumors may be bigger and/or have begun to spread to nearby lymph nodes. When cancer has advanced to the lymph nodes of the mediastinum, a diagnosis of stage III can be determined (the chest area between the lungs). In Stage IV, the cancer has spread to the lining of the lungs or to other organs [1].

Lung cancer (small and non-small cell) is the second-leading cause of cancer in both men and women (excluding skin cancer) in 2020 [2]. This kind of cancer is on the rise in several countries, particularly in Asia, where the rate increased from 56 percent in 2012 to 58 percent in 2018 [3]. In the year of 2020, lung cancer is the top cause of cancer-related mortality with 1.80 million deaths, followed by colon and rectum cancer with 935 thousand deaths, and liver cancer with 850,000 deaths. Lung cancer has killed 4,319 lives in Malaysia, or 2.57 percent of all mortality based on the latest WHO data published in 2020. Malaysia ranks 77th in the world with a death rate of 15.25 per 100,000 population [4].

Malaysia continues to have the lowest 5-year lung cancer survival rate. Symptoms of lung cancer are unusually detected at an early stage, and more than half of lung cancer patients pass away within the first year after diagnosis. Currently, the main causes of lung cancer are unknown. Yet, certain risk factors and symptoms enhance the likelihood of a person developing lung cancer. There are also a few patients with lung cancer who exhibited no symptoms or identified risk factors [5]. Common lung cancer symptoms include persistent coughing, breathing difficulties, bloody coughing, and a sudden decrease in weight. All these symptoms may appear within one month after a lung cancer diagnosis [6].

This research presents a study on the prediction of high-risk symptoms of lung cancer in Malaysia using the fuzzy linear regression method. The primary goal of this study is to determine the highest-risk signs and symptoms of patients with early lung cancer detection to enhance the likelihood of diagnosing malignancy early and decrease lung cancer mortality. It is important for Malaysians to be aware and acknowledge the highest risk symptoms of lung cancer for them to get early treatment at early stages to increase the likelihood of survival. As for the proposed method, numerous researchers have utilized fuzzy modeling to investigate cases in various fields, including medicine, science, and engineering. Fuzzy modeling is typically used to evaluate more complex scenarios and is reliable. Since 1965, when Lotfi A. Zadeh devised the fuzzy set theory, numerous studies have utilized the fuzzy method. Fuzzy linear regression in 1982 is recognized as the basic model for other fuzzy models. The model is conveyed as a dependable method since it does not need any assumptions to ensure the results obtained are applicable to society. Hence, the results will still be accurate even if the data of the sample is small.

The entirety of this article is structured as follows: Part II emphasizes significant studies on current challenges and ways to resolve them, while Section III outlines the research methodologies. Section IV includes the results and Section V presents the discussion. Section VI concludes the paper and proposes future work.

II. Literature Review

A. Lung Cancer in Malaysia

Lung cancer is the primary cause of cancer-related mortality globally and the most common cause of death in Malaysia, with males surpassing females regardless of the tumor's size, location, or dissemination. Lung cancer contributes to around 15.13 percent of cancer deaths [7]. The reported 1-year survival rate is only 35.5 percent however, the relative 5-year survival rate is only 11.0 percent. The survival rate of lung cancer patients in Malaysia at 1 and 5 years is one of the lowest compared to other types of cancer, as shown in Chart A. This survival rate is one of the lowest in the world. The one-year and five-year survival rates are shown per stage in Chart B [8]. Fig. 1 displays Chart A and Chart B.

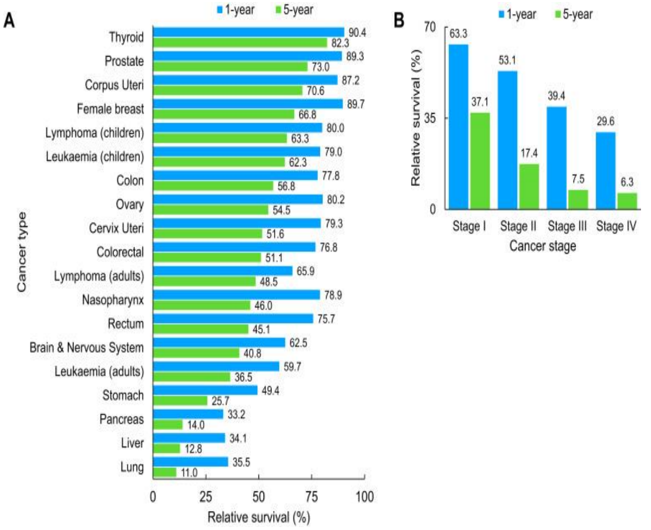

Fig. 1. Relative survival of cancer patients in Malaysia.

In Malaysia, the probability of getting lung cancer is approximately 1 in 60 for males and 1 in 138 for females, with patients often being diagnosed at theto age of 70 or older (range 15 to 90 years). Nevertheless, most cases of lung cancer were not discovered until a very late stage, stage III or stage IV, which is above 90 percent for both sexes. Early-stage disease (I, II, and chosen IIIa) is amenable to curative surgery, which offers the best possibility of long-term cure and disease-free survival [9]. However, most patients (about 75 percent) are diagnosed with advanced cancer (stage III/IV). Despite significant advances in late-stage lung cancer oncological treatment in recent years, survival rates remain poor [10].

There are two distinct diagnostic presentations for lung cancer patients: symptomatic and incidental. Most cases were inadvertently diagnosed through chest X-rays and CT scans. According to the Malaysian Health Technology Assessment Section, a CT scan known as low-dose computed tomography (LDCT) is currently used for lung cancer screening and improved lung cancer diagnosis. Unfortunately, it does not apply to lung cancer patients at high risk. Screening high-risk individuals through screening results is critical since it enhances the likelihood of early cancer detection and decreases lung cancer mortality [11].

While among symptomatic patients, the most often reported complaints that resulted in an imaging referral were the development of a new cough or the worsening of a previously expressed clinical picture suggestive of pneumonia and haemoptysis [12]. It has been proven that cough is the symptom that appears most frequently in lung cancer patients based on the results [13] which frequently reported lung cancer symptoms to include shortness of breath, cough, and anxiety. The study [14] also stressed that fever and cough were the most prominent early symptoms, and respiratory symptoms were prevalent among lung cancer patients. Research [15] reported that cough has the highest number which is 62.0% of patients, followed by chest pain (51.8%) was the most prevalent symptom present at the time of diagnosis. Studies [16] and [17] discovered that haemoptysis had the highest diagnostic value for lung cancer.

B. Background of Fuzzy Linear Regression

Regression analysis is a statistical technique used to determine the cause-and-effect relationship between two variables. Regression analysis is a potent method for comprehending (including forecasting and explaining) the causal factors underlying a population outcome [18]. However, regression models are particularly susceptible to outliers. An outlier is a data point that deviates significantly from most other observations. Variability in measurement may result in an experimental error, whereas an outlier in regression analysis may cause a significant issue. Although data are infrequently linearly separable, regression analysis methods also oversimplify real-world data and issues.

Fuzzy linear regression analysis on the other hand is a significant alternative to conventional regression methods based on statistics. In fuzzy linear regression analysis, a wide variety of fuzzy linear models can be used to approximate a linear dependence based on a set of observations. There are two types of fuzzy regression. The researchers in [19] created 'possibilistic' fuzzy regression, a linear programming method that aims to reduce the fuzziness of a system. The second approach is a fuzzy least-squares method that minimizes the distance between two fuzzy numbers. The approaches are designed to handle fuzzy data to satisfy a particular requirement [20].

In a fuzzy environment, a fuzzy regression model is applied to evaluate the functional relationship between the dependent and independent variables. In the literature, numerous fuzzy regression models and methods for estimating the fuzzy parameters of these models have been developed. The possibilistic approach and fuzzy least squares model are the two most frequent methods for assessing fuzzy regression models [21]. Fuzzy methodology surpasses conventional regression methodology when it is required to predict an outcome variable based on many interrelated factors. Furthermore, it is proved that fuzzy linear regression is more effective relative to simple and multiple linear regression methods [22].

III. Methodology

Fuzzy sets can be used to account for data inaccuracy and ambiguity. For example, rather than just assigning a binary value to a symptom, such as "present" or "absent," a fuzzy set may be used to represent the degree to which a patient exhibits a specific symptom. This would be preferable to the traditional approach of assigning a binary value [23].

The fuzzy linear regression approach is simpler and more transparent to calculate than classical regression but does not significantly differ from classical regression. Furthermore, these results provide support for the concept of fuzzy linear regression prediction, especially when it comes to fuzzy data [24]. The high-risk symptoms of lung cancer can be detected with greater precision by using the fuzzy linear regression method, which provides a better prediction of imprecise data than regression analysis.

Statistical analysis is adaptable and useful to numerous domains, especially the linear regression technique. Several model elements are represented by fuzzy numbers in fuzzy linear regression, which is a kind of regression analysis. It has been demonstrated that fuzzy linear functions are a good strategy for unclear occurrences in linear regression models [25]. The data were analyzed using the statistical software Matlab and Microsoft Excel.

A. Fuzzy Linear Regression (Tanaka, 1982)

To formulate a fuzzy linear regression model, the following were assumed to hold (Tanaka, 1982):

(1) The data can be represented by a fuzzy linear model:

$$\mathrm{Y_e}^* = \mathrm{A_1}^*\mathrm{x}_{\mathrm{e}\,1} + \ldots + \mathrm{A_g}^*\mathrm{x}_{\mathrm{e}\,\mathrm{g}} \triangleq \mathrm{A}^*\mathrm{x_e} \qquad (1)$$

Where,

Fuzzy parameter A_g

The variable of fuzzy parameter x_e

Equation of the fuzzy parameter Y_e^*

$$\mu_{Y_{e^*}}(y) = 1 - \frac{|y_e - x_e^T \alpha|}{\varsigma^T |x_e|} \qquad (2)$$

(2) The degree of the fitting of the estimated fuzzy linear model $Y_e{}^*$ =$A^* x_e$ to the given data Y_e = (y_e , ε_e) was measured by the following index h_e , which maximizes h subject to $Y_e^h \subset Y_e^{*h}$, where:

$$Y_e^h = \{y | \mu_{Ye}(y) \geq h\}$$

$$Y_e^* = \left\{y \,| \mu_{Y_e^*}(y) \geq h\right\} \quad (3)$$

Which are h -level sets. This index h_e is illustrated in Fig. 2. The degree of the fitting of the fuzzy linear model for all data Y_1 , ..., Y_N is defined by $min\ _f\ [h_f\ ]$. Fig. 2 portrays the fitting degree of Y_e^* to a fuzzy data of Y_e.

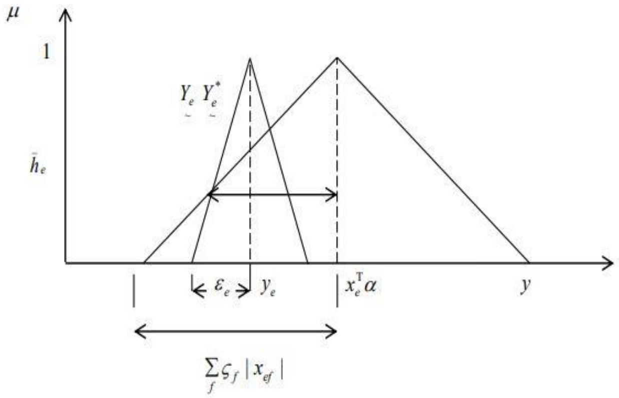

Fig. 2. Degree of fitting of Y_e^* to a given fuzzy data Y_e.

(3) The vagueness of the fuzzy linear model is defined by:

$$\mathrm{JJ} = \varsigma_1 + \ldots + \varsigma_g \quad (4)$$

The problem was elucidated by acquiring fuzzy parameters A* which minimized JJ subject to $\bar{h}$e $\geq H$ for all e , where H was selected by the decision maker as the degree of fit of the fuzzy linear model. The $\bar{h}$e can be acquired by utilizing:

$$\bar{h}\mathrm{e} = 1 - \frac{\left|y_{e\,-}\, x_e^T\, \alpha\right|}{\sum_f \varsigma_f \left|x_{ef}\right| - \varepsilon_e} \qquad (5)$$

Tanaka (1982) model estimated the fuzzy parameter $\boldsymbol{A_e^*} = (\boldsymbol{\alpha_e}, \boldsymbol{\varsigma_e})$, which are the solutions of the following linear programming problem:

$$\min_{\alpha,\varsigma} = \boldsymbol{\varsigma_1} + _{\ldots} + \boldsymbol{\varsigma_g}$$

Subject to $\boldsymbol{\varsigma} \geq \mathbf{0}$ and

$$\alpha^T x_e + (1-H) \sum_f \varsigma_f \left|x_{ef}\right| \geq y_e + (1-H)\varepsilon_e$$

$$-\alpha^T x_e + (1-H) \sum_f \varsigma_f \left|x_{ef}\right| \geq -y_e + (1-H)\varepsilon_e \qquad (6)$$

The best fitting model for the given data may be obtained by solving the conventional linear programming problem in (6). The number of constraints, 2 N, was generally substantially greater than the number of variables, g. As a result, solving the dual problem of (6) was easier than solving the primal problem of (6).

The fuzzy linear regression model (FLRM) can be stated as:

$$Y = A_0\ (\ \alpha_0\ ,\ \varsigma_0\ ) + A\ (\alpha,\ \varsigma)\ \mathrm{x} + \ldots + A\ (\alpha,\ \varsigma)\ x_g \qquad (7)$$

IV. Results

Hideo Tanaka presented a fuzzy approach for linear regression analysis in 1982. There is a fuzzy model in which human estimation and some systems play a role and must deal with a fuzzy structure. Tanaka's fuzzy linear regression was used in the study to estimate the size of lung cancer patients' tumors. In total, 124 patients were used. Gender, ethnic, age, tumor size, cough, haemoptysis, weight loss, appetite loss, chest pain, comorbidities, smoking habit, and stage of cancer were the most key factors. The data were obtained using Microsoft Excel and MATLAB software. H-values ranging from 0.0 to 1.0 were utilised to calculate the center, a_i, and width, c_i, of each fuzzy parameter. a_i is the center of a fuzzy parameter, while c_i represents the parameter's fuzziness (width). The results of this H-value are shown in the Tables I.

TABLE I. FUZZY PARAMETER OF *H*=0.0

Variables	**Fuzzy Parameter**	
	Center a_i	***Width c_i***
Age	0.3332	0.6097
Ethic	3.3148	0
Cough (A_1*)	2.3085	0
Haemoptysis (A_2*)	14.5494	0
Weight loss (A_3*)	5.3752	0
Appetite loss (A_4*)	-6.6669	0
Chest pain (A_5*)	10.6765	0
Smoking habit (A_6*)	-0.0589	0
Comorbidity (A_7*)	-4.8611	0

Table I displays the centre, a_i, and width, c_i values for the fuzzy parameters when H = 0.0. The values of the fuzzy parameter are displayed in Table I; the data was conducted using Matlab code and eleven variables were included. The dependent variable is the size of the tumor, and nine independent variables, lung cancer symptoms. The fuzzy mean tumor size (mm) can be represented by haemoptysis with a fuzzy parameter value of 14.5494. The second highest fuzzy parameter was chest pain equal to 10.6765. The fuzziness of the nine variables reflects the uncertainty of tumor size in millimeters. By this fuzziness parameter, the dispersion can be explained. In this context, the fuzziness of the parameter is J = 0.6097. In addition, the negative nature of A_4*, A_6*, and A_7* is dependent upon the strong correlations between x4, x6, and x7. The tumor size of lung cancer (mm) is inversely proportional to appetite loss, smoking, and comorbidity.

The following is the estimated fuzzy linear regression model for lung cancer patients.

$$\hat{Y} = 0.6097 + (0.3332, 0.6097)\text{ age} + (3.3148, 0)\text{ ethnic} + (2.3085, 0)\text{ cough} + (14.5494, 0)\text{ haemoptysis} + (5.3752, 0)\text{ weight loss} – (6.6669, 0)\text{ loss of appetite} + (10.6765, 0)\text{ chest pain} – (0.0589,0)\text{ smoking} – (4.8611,0)\text{ comorbidity.} \quad (8)$$

TABLE II. FUZZY PARAMETER OF *H*=0.1

Variables	**Fuzzy Parameter**	
	Center a_i	***Width c_i***
Age	0.3336	0.6719
Ethic	3.1988	0
Cough (A_1*)	2.5589	0
Haemoptysis (A_2*)	14.5031	0
Weight loss (A_3*)	5.1294	0
Appetite loss (A_4*)	-6.3314	0
Chest pain (A_5*)	10.4874	0
Smoking habit (A_6*)	-0.0555	0
Comorbidity (A_7*)	-4.7618	0

The following is the estimated fuzzy linear regression model for lung cancer patients:

$$\hat{Y} = 0.6719 + (0.3336, 0.6719)\text{ age} + (3.1988, 0)\text{ ethnic} + (2.5589, 0)\text{ cough} + (14.5031, 0)\text{ haemoptysis} + (5.1294, 0)\text{ weight loss} – (6.3314, 0)\text{ loss of appetite} + (10.4874, 0)\text{ chest pain} – (0.0555,0)\text{ smoking} – (4.7618,0)\text{ comorbidity.} \quad (9)$$

TABLE III. FUZZY PARAMETER OF *H*=0.2

Variables	**Fuzzy Parameter**	
	Center a_i	***Width c_i***
Age	0.3340	0.7498
Ethic	3.0827	0
Cough (A_1*)	2.8094	0
Haemoptysis (A_2*)	14.4567	0
Weight loss (A_3*)	2.5344	0
Appetite loss (A_4*)	-3.6466	0
Chest pain (A_5*)	10.2983	0
Smoking habit (A_6*)	-0.0521	0
Comorbidity (A_7*)	-4.7618	0

The following is the estimated fuzzy linear regression model for lung cancer patients:

$$\hat{Y} = 0.7498 + (0.3340, 0.7498)\text{ age} + (3.0827, 0)\text{ ethnic} + (2.8094, 0)\text{ cough} + (14.4567, 0)\text{ haemoptysis} + (2.5344, 0)\text{ weight loss} – (3.6466, 0)\text{ loss of appetite} + (10.2983, 0)\text{ chest pain} – (0.0521,0)\text{ smoking} – (4.6624,0)\text{ comorbidity.} \quad (10)$$

TABLE IV. FUZZY PARAMETER OF *H*=0.3

Variables	**Fuzzy Parameter**	
	Center a_i	***Width c_i***
Age	0.3344	0.8498
Ethic	2.9667	0
Cough (A_1*)	3.0599	0
Haemoptysis (A_2*)	14.4104	0
Weight loss (A_3*)	4.6379	0
Appetite loss (A_4*)	-5.6603	0

Variables	Fuzzy Parameter	
	Center a_i	**Width c_i**
Chest pain (A_5*)	10.1093	0
Smoking habit (A_6*)	-0.0487	0
Comorbidity (A_7*)	-4.5630	0

The following is the estimated fuzzy linear regression model for lung cancer patients:

$$\hat{Y} = 0.8498 + (0.3344, 0.8498) \text{ age} + (2.9667, 0) \text{ ethnic} + (3.0599, 0) \text{ cough} + (14.4104, 0) \text{ haemoptysis} + (4.6379, 0) \text{ weight loss} - (5.6603, 0) \text{ loss of appetite} + (10.1093, 0) \text{ chest pain} - (0.0487,0) \text{ smoking} - (4.5630,0) \text{ comorbidity.} \quad (11)$$

TABLE V. FUZZY PARAMETER OF *H*=0.4

Variables	Fuzzy Parameter	
	Center a_i	**Width c_i**
Age	0.3348	0.9833
Ethic	2.8506	0
Cough (A_1*)	3.3103	0
Haemoptysis (A_2*)	14.3640	0
Weight loss (A_3*)	4.3922	0
Appetite loss (A_4*)	-5.3248	0
Chest pain (A_5*)	9.9202	0
Smoking habit (A_6*)	-0.0453	0
Comorbidity (A_7*)	-4.4636	0

The following is the estimated fuzzy linear regression model for lung cancer patients:

$$\hat{Y} = 0.9833 + (0.3348, 0.9833) \text{ age} + (2.8506, 0) \text{ ethnic} + (3.3103, 0) \text{ cough} + (14.3640, 0) \text{ haemoptysis} + (4.3922, 0) \text{ weight loss} - (5.3248, 0) \text{ loss of appetite} + (9.9202, 0) \text{ chest pain} - (0.0453,0) \text{ smoking} - (4.4636,0) \text{ comorbidity.} \quad (12)$$

TABLE VI. FUZZY PARAMETER OF *H*=0.5

Variables	Fuzzy Parameter	
	Center a_i	**Width c_i**
Age	0.3317	1.1718
Ethic	2.7910	0
Cough (A_1*)	3.4661	0
Haemoptysis (A_2*)	14.3992	0
Weight loss (A_3*)	2.3009	0
Appetite loss (A_4*)	-3.1698	0
Chest pain (A_5*)	9.8490	0
Smoking habit (A_6*)	-0.0424	0
Comorbidity (A_7*)	-4.3946	0

The following is the estimated fuzzy linear regression model for lung cancer patients:

$$\hat{Y} = 1.1718 + (0.3317, 1.1718) \text{ age} + (2.7910, 0) \text{ ethnic} + (3.4661, 0) \text{ cough} + (14.3992, 0) \text{ haemoptysis} + (2.3009, 0) \text{ weight loss} - (3.1698, 0) \text{ loss of appetite} + (9.8490, 0) \text{ chest pain} - (0.0424,0) \text{ smoking} - (4.3946,0) \text{ comorbidity.} \quad (13)$$

TABLE VII. FUZZY PARAMETER OF *H*=0.6

Variables	Fuzzy Parameter	
	Center a_i	**Width c_i**
Age	0.3278	1.4551
Ethic	2.7459	0
Cough (A_1*)	3.5973	0
Haemoptysis (A_2*)	14.4555	0
Weight loss (A_3*)	2.1314	0
Appetite loss (A_4*)	-2.9434	0
Chest pain (A_5*)	9.8084	0
Smoking habit (A_6*)	-0.0397	0
Comorbidity (A_7*)	-4.3334	0

The following is the estimated fuzzy linear regression model for lung cancer patients:

$$\hat{Y} = 1.4551 + (0.3278, 1.4551) \text{ age} + (2.7459, 0) \text{ ethnic} + (3.5973, 0) \text{ cough} + (14.4555, 0) \text{ haemoptysis} + (2.1314, 0) \text{ weight loss} - (2.9434, 0) \text{ loss of appetite} + (9.8084, 0) \text{ chest pain} - (0.0397,0) \text{ smoking} - (4.3334,0) \text{ comorbidity.} \quad (14)$$

TABLE VIII. FUZZY PARAMETER OF *H*=0.7

Variables	Fuzzy Parameter	
	Center a_i	**Width c_i**
Age	0.3238	1.9273
Ethic	2.7009	0
Cough (A_1*)	3.7285	0
Haemoptysis (A_2*)	14.5119	0
Weight loss (A_3*)	1.9619	0
Appetite loss (A_4*)	-2.7169	0
Chest pain (A_5*)	9.7678	0
Smoking habit (A_6*)	-0.0370	0
Comorbidity (A_7*)	-4.2723	0

The following is the estimated fuzzy linear regression model for lung cancer patients:

$$\hat{Y} = 1.9273 + (0.3238, 1.9273) \text{ age} + (2.7009, 0) \text{ ethnic} + (3.7285, 0) \text{ cough} + (14.5119, 0) \text{ haemoptysis} + (1.9619, 0) \text{ weight loss} - (2.7169, 0) \text{ loss of appetite} + (9.7678, 0) \text{ chest pain} - (0.0370,0) \text{ smoking} - (4.2723,0) \text{ comorbidity.} \quad (15)$$

TABLE IX. FUZZY PARAMETER OF *H*=0.8

Variables	Fuzzy Parameter	
	Center a_i	***Width c_i***
Age	0.3198	2.8718
Ethic	2.6559	0
Cough (A_1*)	3.8598	0
Haemoptysis (A_2*)	14.5682	0
Weight loss (A_3*)	4.1194	0
Appetite loss (A_4*)	-4.8174	0
Chest pain (A_5*)	9.7272	0
Smoking habit (A_6*)	-0.0315	0
Comorbidity (A_7*)	-4.2111	0

The following is the estimated fuzzy linear regression model for lung cancer patients:

$$\hat{Y} = 2.8718 + (0.3198, 2.8718)\text{ age} + (2.6559, 0)\text{ ethnic} + (3.8598, 0)\text{ cough} + (14.5682, 0)\text{ haemoptysis} + (4.1194, 0)\text{ weight loss} - (4.8174, 0)\text{ loss of appetite} + (9.7272, 0)\text{ chest pain} - (0.0315,0)\text{ smoking} - (4.2111,0)\text{ comorbidity.} \quad (16)$$

TABLE X. FUZZY PARAMETER OF *H*=0.9

Variables	Fuzzy Parameter	
	Center a_i	***Width c_i***
Age	0.3158	5.7051
Ethic	2.6109	0
Cough (A_1*)	3.9910	0
Haemoptysis (A_2*)	14.6245	0
Weight loss (A_3*)	4.0609	0
Appetite loss (A_4*)	-4.7019	0
Chest pain (A_5*)	9.6866	0
Smoking habit (A_6*)	-0.0315	0
Comorbidity (A_7*)	-4.1500	0

The following is the estimated fuzzy linear regression model for lung cancer patients:

$$\hat{Y} = 5.7051 + (0.3158, 5.7051)\text{ age} + (2.6109, 0)\text{ ethnic} + (3.9910, 0)\text{ cough} + (14.6245, 0)\text{ haemoptysis} + (4.0609, 0)\text{ weight loss} - (4.7019, 0)\text{ loss of appetite} + (9.6866, 0)\text{ chest pain} - (0.0315,0)\text{ smoking} - (4.1500,0)\text{ comorbidity.} \quad (17)$$

TABLE XI. FUZZY PARAMETER OF *H*=1.0

Variables	Fuzzy Parameter	
	Center a_i	***Width c_i***
Age	0.3128	5.9810
Ethic	2.5509	0
Cough (A_1*)	4.1223	0
Haemoptysis (A_2*)	14.6450	0
Weight loss (A_3*)	4.0024	0
Appetite loss (A_4*)	-4.6237	0
Chest pain (A_5*)	9.6400	0
Smoking habit (A_6*)	-0.0315	0
Comorbidity (A_7*)	-4.1034	0

Table II to Table XI show the centre, a_i, and width, c_i values for the fuzzy parameters when H = 0.1 until H = 1.0. The following is the estimated fuzzy linear regression model for lung cancer patients:

$$\hat{Y} = 5.9810 + (0.3128, 5.9810)\text{ age} + (2.5509, 0)\text{ ethnic} + (4.1223, 0)\text{ cough} + (14.6450, 0)\text{ haemoptysis} + (4.0024, 0)\text{ weight loss} - (4.6237, 0)\text{ loss of appetite} +(9.6400, 0)\text{ chest pain} - (0.0315,0)\text{ smoking} - (4.1034,0)\text{ comorbidity.} \quad (18)$$

1) Measuring mean square error (MSE): Table XII displays the mean square error (MSE) values for those H-values. The observed Y is determined by the responses of 124 patients with lung cancer. The *H*-value with the smallest MSE is 0.0. Since the MSE value of the *H*-value of 0.0 is the lowest when compared to other values, it has been concluded that this model is the most suited and effective model for predicting the high-risk symptoms of lung cancer.

TABLE XII. MEAN SQUARE ERROR VALUES

MSE Values	
H-values	***Mean Square Error***
0.0	1.455
0.1	1.467
0.2	1.481
0.3	1.497
0.4	1.517
0.5	1.549
0.6	1.592
0.7	1.657
0.8	1.774
0.9	2.101
1.0	2.138

2) Measuring root mean square error (RMSE): Table XIII shows the root mean square error, which is computed by calculating the square root of the total mean square error to achieve the least error value. The models were tested using mean square error. H-values of 0.0 and 1.0 have RMSE values of 1.206 and 1.462, respectively. The fuzzy linear regression model with H-value of 0.0 proves to be the most precise model for predicting the high-risk symptoms reported by lung cancer patients at Hospital Al-Sultan Abdullah (UiTM Hospital), given that its RMSE is the lowest among the other models.

TABLE XIII. ROOT MEAN SQUARE ERROR VALUES

RMSE Values	
H-values	***Root mean square error***
0.0	1.206
0.1	1.211
0.2	1.217
0.3	1.224
0.4	1.232
0.5	1.244
0.6	1.262
0.7	1.287
0.8	1.332
0.9	1.450
1.0	1.462

V. DISCUSSION

Fuzzy linear regression with an H-value of 0.0 is the best model for predicting high-risk lung cancer symptoms in patients at Al-Sultan Abdullah Hospital (UiTM Hospital). Fuzzy linear regression with an H-value of 0.0 has lowest mean square error (MSE) and root mean square error (RMSE) values compared to other H-values. H-value of 0.0 produced MSE and RMSE values of 1.455 and 1.206, while an H-value of 1.0 yielded MSE and RMSE values of 1.784 and 1.336, respectively. The optimal model has been proved to be the fuzzy linear regression of H-value with 0.0 as it has the smallest measurement error. The summary values of MSE and RMSE are displayed in Table XIV.

TABLE XIV. MSE AND RMSE VALUES OF THE MODELS

Summary of MSE and RMSE Values		
H-values	***MSE***	***RMSE***
0.0	1.455	1.206
0.1	1.467	1.211
0.2	1.481	1.217
0.3	1.497	1.224
0.4	1.517	1.232
0.5	1.549	1.244
0.6	1.592	1.262
0.7	1.657	1.287
0.8	1.774	1.332
0.9	2.101	1.450
1.0	2.138	1.462

The best parameter for fuzzy linear regression was discovered using the values of mean square error and root mean square error. This study found that haemoptysis was the most impactful symptom in diagnosing lung cancer high-risk symptoms, as it has the highest fuzzy mean parameter in the model with an H-value of 0.0 and a value of 14.5494, as shown in Table I. The findings similar to the [26] stated that individuals diagnosed with lung cancer showed a significantly greater prevalence of persistent haemoptysis. [27] and [28] also emphasized that haemoptysis is the most prevalent cause of lung cancer across all age groups. Chest pain is the second highest risk symptom of lung cancer as it has the value of fuzzy mean parameter of 10.6765 based on Table I. The results are also akin to the study by [29] found that the frequency of haemoptysis, cough, and chest pain was significantly higher than in other samples in all stages. According to [30] chest pain was the top five major complaints when it comes to lung cancer symptoms. Age, ethnicity, cough and weight loss are the other variables that are closely related to the high-risk symptoms of lung cancer as it has positive values of fuzzy mean parameter. In addition, the high-risk symptoms of lung cancer are inversely proportional to the female gender, Chinese ethnicity and other ethnicities, appetite loss, smoking habit, and the presence of other diseases (comorbidity) as it has negative values of fuzzy mean parameters. The lowest mean square error is 1.455, and the least root mean square error is 1.206.

The purpose of this study was to predict the high-risk symptoms of lung cancer in the early stage to initiate preventative measures for lung cancer patients. It was determined that haemoptysis and chest pain are high-risk symptoms for lung cancer patients at Hospital Al-Sultan Abdullah (UiTM Hospital). However, most of the patient data collected from Al-Sultan Abdullah hospital (UiTM Hospital) was from patients with advanced lung cancer (stages 3 and 4). It is difficult to detect symptoms for the earlier stage of lung cancer due to the tumor size in stages I and II is smaller in size since the small tumors convey less texture and shape information than those larger tumors in late stages [31]. It is stated that the bigger the diameter of the tumor size, the more advanced the stage of lung cancer, and the symptoms of lung cancer will begin to appear one by one such as haemoptysis and chest pain. Even though the symptoms are recognized at a late stage, the doctors or patients can still take the initiative or precautions at earlier stages for more particular symptoms as revealed by the results rather than any random symptoms.

VI. CONCLUSION

The aim of this study was to determine high-risk lung cancer symptoms in order to initiate preventative actions. It was concluded that haemoptysis and chest pain are high-risk symptoms for lung cancer patients at Hospital Al-Sultan Abdullah (UiTM Hospital). Both high-risk symptoms can be presented to medical doctors and nurses at the UiTM Hospital so that they can apply them to patients at an early stage. Extreme weight loss, loss of appetite, and comorbidity are the further lung cancer symptoms. Fuzzy linear regression with *H*-value of 0.0 is the best model for predicting the high-risk symptoms of lung cancer in patients at Hospital Al-Sultan Abdullah (UiTM Hospital) as it has the least measurement error, with mean square error (MSE) and root mean square error (RMSE) values of 1.45 and 1.206, respectively.

In future studies, other researchers should resolve the issue of determining the stages of lung cancer among patients at Selangor's general hospitals. The study should be expanded to include other public hospitals in each state of Malaysia. In that scenario, the lung cancer study may be thoroughly explored in Malaysia and other countries.

ACKNOWLEDGMENT

This research was supported by Ministry of Higher Education (MOHE) through Fundamental Research Grant Scheme (FRGS/1/2021/STG06/UTHM/03/1).

REFERENCES

[1] American Lung Association. “Lung cancer staging”, 2022.

[2] American Cancer Society, “Information and Resources about for Cancer: Breast, Colon, Lung, Prostate, Skin”, 2023.

[3] R. Pakzad, A. Mohammadian-Hafshejani, M. Ghoncheh, I. Pakzad, and H. Salehiniya, “The incidence and mortality of lung cancer and their relationship to development in Asia,” Translational lung cancer research, 4(6), 763–774, 2015.

[4] World Health Organization, Malaysia Sources: Globocan 2020, International Agency for Research of Cancer, 1-2, 2020.

[5] M. Mustafa, AR. J. Azizi, A. Nazirah, A. M. Sharifa, and S. A. Abbas, “Lung Cancer: Risk Factors, Management, And Prognosis,” IOSR Journal of Dental and Medical Sciences, Vol. 15, No.10, p. 94-101, 2016.

[6] R. Gasparri, M. Santonico, C. Valentini, and G. Sedda. (2016). “Volatile signature for the early diagnosis of lung cancer,” Journal of Breath Research, p. 1-7.

[7] J. C. Alcantud, G. Varela, B. S. Buitrago, G. S. Garcia, and M. F. Jimenez, “Analysis of survival for lung cancer resections cases with Fuzzy and soft set theory in surgical decision making” PLoS ONE, Vol. 14, No. 6, p.1–17, 2019.

[8] National Cancer Institute, Malaysia National Cancer Registry Report (MNCR) 2012-2016, p. 100, 2019. [Putrajaya: Ministry of Health Malaysia].

[9] A. Sachithanandan, and B. Badmanaban, “Screening for Lung cancer in Malaysia: Are we there yet?,” Medical Journal of Malaysia, Vol. 67 No. 1,p. 3–6, 2012.

[10] S. Blandin Knight, P. Crosbie, H. Balata, J. Chudziak, T. Hussell, and C. Dive, “Progress and prospects of early detection in lung cancer,” *Open Biology*, Vol. *7* No. 9, 170070, 2017.

[11] M. Wille, A. Dirksen, H. Ashraf, Z. Saghir, K. Bach, and J. Brodersen, “Results of the Randomized Danish Lung Cancer Screening Trial with Focus on High-Risk Profiling, “ American Journal Of Respiratory And Critical Care Medicine, 2016.

[12] S. Quadrelli., G. Lyons, H. Colt, D. Chimondeguy, and A. Buero, “Clinical Characteristics and Prognosis of Incidentally Detected Lung Cancers,” International Journal Of Surgical Oncology, 2015.

[13] M.S. Whisenant, L.A. Williams, A.G. Gonzalez, and T. Mendoza, “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 ? Vol. 16 No.10, 2022.

[14] L. Nie, K. Dai, J. Wu, X. Zhou., J. Hu, C. Zhang, Y. Zhan, Y. Song, W. Fan, Z. Hu, H. Yang, Q. Yang, D. Wu, F. Li, D. Li, and R. Nie, “Clinical characteristics and risk factors for in-hospital mortality of lung cancer patients with COVID-19: A multicenter, retrospective, cohort study,” Thoracic Cancer, Vol.12 No.1, p.57–65, 2021.

[15] Galvez, M., Rossana, N., Joseph, R., Katia, A. P., Raul, R., & Luis, M, “Lung Cancer in the Young,” 2019.

[16] Garg, A., Jain, V. K., Mishra, M., Maan, L., Jain, G., & Bhardwaj, G. “ To study the Prevalence and Pattern of Haemoptysis in Histopathologically proven cases of Lung cancer and its relation with various Histopathological types of malignancy,” Vol 18, No.6, p.39-41, 2019.

[17] Okoli, G. N., Kostopoulou, O., & Delaney, B. C. “Is symptom-based diagnosis of lung cancer possible? A systematic review and meta-analysis of symptomatic lung cancer prior to diagnosis for comparison with real-time data from routine general practice,” PLOS ONE, Vol.13, No.11, p.1-17, 2018.

[18] Jihye, J. “The Strengths and Limitations of the Statistical Modeling of Complex Social Phenomenon: Focusing on SEM, Path Analysis, or Multiple Regression Models,” International Journal Of Economics And Management Engineering, Vol.9, No.5, p.9,2021.

[19] Tanaka, H., Uejima, S. and Asai, K. “Linear Regression Analysis with Fuzzy Model,” IEEE Transactions On Systems, Man and Cybernetics, SMC-12, p.903-907, 1982.

[20] Khan, U., & Valeo, C. “A new fuzzy linear regression approach for dissolved oxygen prediction,” Hydrological Sciences Journal, Vol.60, No.6, p.1096-1119, 2015.

[21] Denoda, L., Casas Cardoso, G., Luis Morales Martínez, J., González Rodríguez, E., & Rodríguez Corvea, L. “Fuzzy linear regression models: a medical application,” 2014.

[22] Pandit, P., Dey, P., & Krishnamurthy, K. N. “Comparative Assessment of Multiple Linear Regression and Fuzzy Linear Regression Models,” SN Computer Science, Vol.2, No.2, p.1–8, 2021.

[23] Thomas, L. L., Goni, I., & Emeje, G. D. “Fuzzy Models Applied to Medical Diagnosis: A Systematic Review,” Advances in Networks, Vol.7, No.2, p.45–50, 2019.

[24] Al-Sabri, E. H. “The fuzzy linear regression,” Asia Pacific Journal of Mathematics, Vol.7, No.7, 2020.

[25] Munawar, Z., Ahmad, F., Awadh Alanazi, S., Nisar, K. S., Khalid, M., Anwar, M., & Murtaza, K. “Predicting the prevalence of lung cancer using feature transformation techniques,” Egyptian Informatics Journal, Vol. 23, No. 4, p.109-120, 2022.

[26] Arooj, P., Bredin, E., Henry, M. T., Khan, K. A., Plant, B. J., Murphy, D. M., & Kennedy, M. P. “Bronchoscopy in the investigation of outpatients with hemoptysis at a lung cancer clinic,” Respiratory Medicine, Vol.139, p.1–5, 2018.

[27] Lasake, I. B., Idayu, R., Mat, B., Binti, N., & Marzuki, M. “Recurrent Haemoptysis in Non-Small Cell Lung Cancer Patient,”. Vol.2, No.1, p.18–19, 2020.

[28] Bankar, A., Padamwar, K., & Jahagirdar, A. “Symptom analysis using a machine learning approach for early stage lung cancer;” Proceedings of the 3rd International Conference on Intelligent Sustainable Systems, ICISS 2020, p.246–250, 2020.

[29] 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. “Lung cancer symptoms at diagnosis: Results of a nationwide registry study,” ESMO Open, Vol.5, No.6, p.1–7, 2020.

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

[31] Chaddad, A., Desrosiers, C., Toews, M., & Abdulkarim, B. “Predicting survival time of lung cancer patients using radiomic analysis,” Oncotarget, Vol.8, No.61, p.104393-104407, 2017.