OPEN ACCESS
Indo-JC Indonesia Journal of Computing
ISSN 2460-9056
socj.telkomuniversity.ac.id/indojc

Ind. Journal on Computing
Vol. 7, Issue. 2, August 2022. pp. 1-14
doi:10.34818/indojc.2022.7.2.616

Lung Cancer Prediction Model using Logistic Linear Regression with Imbalanced Dataset

Priscilia Lovita Paelongan [#1], Irma Palupi [#2]

 # School of Computing, Telkom University
Bandung, West Java, Indonesia

[1] lovitapriscilia@gmail.com
[2] irmapalupi@telkomuniversity.ac.id

Abstract

Cancer is one of the leading causes of death worldwide. Cancer cases in Indonesia have now reached 4.8 million in 2018. Most cases are breast, cervix, and lung. Furthermore, we need to note that 43 percent of these cancer cases are preventable. This study uses a linear logistics regression model. Linear logistic regression models can be used for categoric datasets. The appropriate model is obtained after parameter assessment, test the significance of each affecting attribute, and test the suitability of the model. This is done to obtain prediction models and risk factors at the level of correlation of disease size. This method is relatively easy and conceptually practical, so it is possible to apply it to diagnose early symptoms of lung cancer. The results include a linear logistics regression model for early prediction of lung cancer patients based on symptoms, habits, and history of health diseases to see the likelihood that someone with a certain level of risk could have lung cancer. The factors that affect a person with lung cancer are difficulty swallowing, coughing, chronic diseases, fatigue, and age.

Keywords: Lung Cancer, Prediction Model, Logistic Linear Regression, Machine Learning

Abstrak

Kanker merupakan salah satu penyebab kematian utama di seluruh dunia. Kasus kanker di Indonesia kini telah mencapai 4,8 juta pada 2018. Kasus terbanyak adalah payudara, serviks, dan paru. Dan perlu kita catat bahwa 43 persen kasus kanker ini dapat dicegah. Penelitian ini menggunakan model regresi logistik linier. Model regresi logistik linier dapat digunakan untuk dataset yang bersifat kategorik. Model yang sesuai diperoleh setelah dilakukan penaksiran parameter, uji signifikansi setiap atribut yang mempengaruhi, dan uji kecocokan model. Hal tersebut dilakukan untuk mendapatkan model prediksi dan faktor risiko pada tingkat korelasinya dari ukuran penyakit. Metode ini cukup mudah dan praktis secara konseptual, sehingga mungkin untuk diterapkan dalam diagnosa gejala awal kanker paru-paru. Hasil yang didapatkan yaitu model regresi logistik linier untuk prediksi dini penderita kanker paru-paru berdasarkan gejala, kebiasaan, dan riwayat penyakit kesehatan untuk melihat kemungkinan seseorang dengan tingkat resiko tertentu bisa menderita kanker paru-paru. Adapun faktor-faktor yang mempengaruhi seseorang menderita kanker paru-paru yaitu kesulitan menelan, batuk, penyakit kronis, kelelahan, dan umur.

Kata Kunci: Kanker Paru, Model Prediksi, Regresi Logistik Linier, Pembelajaran Mesin

Received on April, 2022. Accepted on July, 2022

I. Introduction

Tumors are conditions where cell growth is abnormal and thus forms a lesion or, in most cases, lumps in the body. A tumor is divided into two, namely benign tumors and malignant tumors. Malignant tumors or known as cancers, such as lung cancer and breast cancer. The study will focus on lung cancer because lung cancer is the leading cause of cancer-related death globally, with mortality rates exceeding prostate cancer, breast cancer, and cervical cancer. In Indonesia, lung cancer is the leading cause of death among men, and more than 70% of cancer cases are diagnosed in the advanced stage [1].

Data from GLOBOCAN 2020 showed that among men and women, the distribution of the most common cancer incidence was lung (19.2%), breast (10.8%), stomach (8.6%), liver (6, 9%) and colon (6.0%). The main causes of cancer death among women and men are lung cancer (19.2%), liver (10.5%), stomach (9.9%), esophageal (7.5%) and breast (6.0%). Lung cancer is also the leading cause of cancer death in Indonesia. Every hour, four patients are diagnosed with lung cancer and three patients die of lung cancer in Indonesia. The incidence and mortality rate of lung cancer in Indonesia is higher than the average in Asia and the world [2].

Lung cancer can be caused by a variety of factors, including cigarette smoke and genetic changes. Smoking is the leading cause of lung cancer in 80-90% of lung cancer cases, even though only 10-15% of smokers have lung cancer [3]. Cancer has high rates of morbidity and mortality. This is a significant threat to human life. Although there are many prevention methods and treatment of tumors, ranging from primary to secondary levels, it is still not enough. Studies have shown that advanced stage lung cancer has a lower overall survival compared to earlier stages. Thus, screening and early detection in primary care health facilities are crucial for lung cancer. Therefore, early detection, early diagnosis, and early treatment are still the main steps to improving survival rates and reducing cancer patient mortality rates [4].

Symptoms do get usually occur until the cancer is advanced, and may include persistent cough, sputum streaked with blood, chest pain, voice change, worsening shortness of breath, and recurrent pneumonia or bronchitis [5]. And also, other symptoms such as chronic cough, hemoptysis (coughing up blood), hoarseness, cachexia (wasting syndrome; weight loss, muscle mass, and fat composition related to cancer), wheezing, chest pain, and clubbing finger [2].

Thus, the formation of tumor prediction models is crucial. Currently, the formation of a predictive model to be used is logistic regression. So, we've discovered an approach that can be used to detect cancer in its early stage with machine learning techniques. We'll use logistic regression to classify the datasets in this work. We use logistic regression for this research because it is the generalized form of linear regression. Primarily, it is worn for computing binary or multi-class dependent variables. Because the reply variable is discrete, it cannot be modelled directly by linear regression. For building a model, it forecast the odds of its event instead of forecasting the point estimate of the occurrence. In two class problem when the result of odds greater than 50% then the class is designated by assigned value 1 otherwise it is 0. However, it is a very ably accepted kit, it allows that the reply variable is linear in the coefficients of the forecast variables. Then, using the experience of data analysis the experimenter must choose the original inputs and decide their functional relationship to the reply to variable [6].

II. Literature Review

A. Related Work

Previous research conducted by Shuting Shen, Ziqiang Fan, and Qi Guo in 2017 [4], lung cancer became a prime use example of logistic regression from analysis of relevant data. The prediction models used are logistic regression and BP neural network as early detection of lung tumors. This process will use the R language to

read and analyze lung cancer statistical data and build logistic regression models for the exploration of risk factors and predictions of pain probabilities.

Besides, risk components obtained based on logistic regression methods will be incorporated into BP neural networks to analyze sensitivity, find more accurate risk factor components, and correlate those factors with tumor pathogenesis. The accuracy of the above model is then verified by designing a tumor prediction model based on a logistic regression model and analysis of tumor risk factor sensitivity based on BP neural networks. The results obtained are three factors that affect lung tumors, namely age, smoking habits, and a history of respiratory diseases with sensitivity as follows: 0.17, 0.48, 0.69. And obtained model design based on logistic regression:

$$\text{logit}(P) = -5.67 + 3.57 \text{ x Age} + 3.20 \text{ x Smoking History} + 1.85 \text{ x History of Respiratory Diseases}$$

Research conducted by Raghavendra Patil G E, Sinchana C G, Tejashwini P, Tejaswini K N, Veena Vittal Ganiga in 2020 [5], using Logistic Regression which will be used to classify the datasets. The aim of this project work is to develop a lung cancer prediction system using logistic regression. First of all, the dataset were partitioned into two parts for training and testing. Then the system is trained using different training dataset using logistic regression such that the system can predict lung cancer in the patients more accurately. Then the system is tested using testing dataset for the accurate result. So, the result obtained is training accuracy = 96% and testing accuracy = 84%.

Research conducted by Animesh Hazra, Nanigopal Bera, Avijit Mandal in 2017 [6], this research inspects the accomplishments of Support Vector Machine (SVM) and Logistic Regression (LR) algorithms in predicting the survival rate of lung cancer patients and compares the effectiveness of these two algorithms through accuracy, precision, recall, F1 score and confusion matrix. In this study data cleaning, feature selection, splitting and classification techniques have been applied for predicting survivability of lung cancer as accurately as possible. This project reveals that logistic regression classifier gives the topmost accuracy of 77.40% compared to support vector machine classifier which gives 76.20% accuracy. Also, the logistic regression classifier gives maximum classification accuracy concerning every different classifier.

Research conducted by Dr. M. Kasthuri dan M. Riyana Jency in 2020 [7], classified about various Machine Learning techniques to predict the lung cancer by using dataset. The accuracy, Precision, Recall and F-Measure are calculated for Support Vector Machine, Naive Bayes, K nearest neighbour, Logistic Regression techniques. The result shows the Support Vector Machine algorithm give the best accuracy of 82.25%.

Like previous studies, this study will also use logistic regression to predict lung cancer. But what makes the difference is that this study uses statistical methods and the dataset used is an imbalanced dataset, so care must be taken in forming the model, especially in the splitting stage of the dataset to train the model and evaluate/test the data. This research also not only provide the accuracy of the best model but also provide the mathematical model of the best model.

B. Cancer

Cancer is a disease that arises from the abnormal growth of body tissue cells that turn into cancer cells where cancer cells can be found in tumor growth that infiltrate the surrounding tissues. The tumor is a condition where abnormal growth of body cells occurs. Each of the cells that make up the human body's tissues contains genes that function in controlling the growth, development, or improvement in the body. During human life, there are times when some cells have to die, divide themselves or turn into certain forms. Some conditions cause the process to be disrupted and trigger old cells not to die when it is time. The non-dead cells will eventually accumulate and unite with the new cells that will form. The cell will then form a mass in which the pile is then called a tumor [8].

There are two categories of tumors, namely benign tumors and malignant tumors. A benign tumor is a collection of cells that grow in only one part of the body. Also, this type of tumor generally does not spread or attack other parts of the body. Instead, malignant tumors are piles of cells that can attack surrounding tissues throughout the body. Malignant tumors can enter blood vessels or into other parts of the human body. The name also knows this type of tumor of cancer [1].

C. Logistic Regression

Logistic regression is one of the multivariate analyses, which is useful for predicting variable dependents based on independent variables. Logistic regression is a type of regression that connects between one or more independent variables with dependent variables in the form of categories; usually 0 and 1.

Logistic regression is a regression model used when the response variable is qualitative. A simple logistic regression model is a logistic regression model for one X-independent variable with a dichotomy Y-dependent variable. The variable value Y = 1 represents the presence of a characteristic and Y = 0 represents the absence of a characteristic [9].

Binary logistics regression is a statistical analysis technique used to analyze the relationship between one or more free changers and binary or dichotomous response changers [10]. Free change on logistic regression can be either a categorical scale change or a continuous-scale changer while the response change is a categorical scale changer. The free modifier is indicated by the vector x' =(x_1,x_2,...,x_p) and the Y response modifier, where Y has two possible values of 0 and 1. The Y response change follows the Bernoulli distribution with the probability function:

$$f(Y = y) = \pi^{y} (1 - \pi)^{1-y} \quad (1)$$

where π is the probability for $y = 1$ occurs.

If the Y response change amounts to n, the chances of each event are the same, and each event is free of each other with other events, then the Y response change will follow the Binomial spread. In the logit regression model, a connecting function is required that corresponds to the logit regression model. Logit transformation as a function of $\pi(x)$ is stated as follows [11] :

$$\text{logit } [\pi(x)] = g(x)=\ln\left[\frac{\pi(x)}{1-\pi(x)}\right] \quad (2)$$

with linear predictor:
$g(x) = \beta_0+\beta_1x_1+\beta_2x_2+...+\beta_px_p$
p = number of free modifiers

The logistic regression model formed by $E(Y=1\cdot x)$ as $\pi(x)$ is as follows:

$$\pi(x) = \left[\frac{\exp(g(x))}{1+\exp(g(x))}\right] \quad (3)$$

The logit transformation aims to create a linear function from its parameters. The g(x) is a linear function that the variables are defined in range $(-\infty. \infty)$ [9].

III. RESEARCH METHOD

A. System Flowchart

The design of the system built and implemented in this study is the design of a prediction model using a static model (logistic regression). The flowchart can be seen in Figure 1.

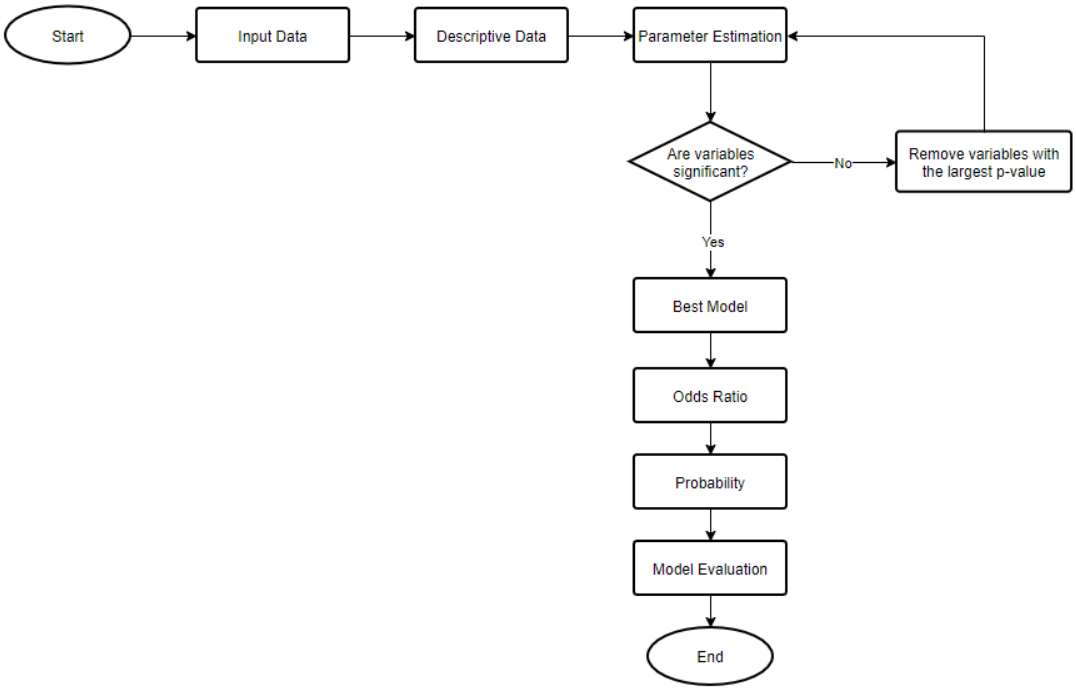

Fig. 1. Flowchart

B. Input Data and Descriptive Data

The input is the data of cancer patients as variables (Y) and factors that affect it as variables (X).

The description of the number of cancer patients performed at this stage aims to analyze the data used by analyzing the number of data samples, the number of cancer survivors, and those who do not have cancer.

C. Parameter Estimation and Significance Test

To get an analysis, the first step is to look for the estimated value of the parameter $\beta_0,\beta_1,\beta_2,\ldots\ \beta_p$ to obtain a binary logistics regression model. At this stage, the process of elimination will be done by eliminating variables whose significance value is more than 0.05. Test the significance of independent variables (X) that affect dependent variables. The process will continue until the model formed does not contain insignificant variables.

D. Best Model

Determining the best model can be done by using the equation formula of the logistic regression model, such as:

$$\text{logit } [\pi(x)] = g(x)=\ln\left[\frac{\pi(x)}{1-\pi(x)}\right] \quad (4)$$

E. Odds Ratio

Determining the Odds Ratio value aims to find out how many risk factors compare to the incidence of cancer sufferers. The results were then compared between risk factors and non-risky factors.

F. Probability

To know the chances of incidence of cancer patients can be computed by using:

$$\text{Probability} = \frac{1}{1+\exp[-g(x_i)]} \tag{5}$$

G. Model Evaluation

Evaluation of binary logistics regression models is based on calibration and discrimination. The obtained calibration value considered to be good when the p-value is greater than 0.05. The value of discrimination is obtained from the Receiver Operating Characteristic (ROC) test to obtain an Area Under the Curve (AUC) value. The AUC value is between 50% to 100%. The closer resulting value to 100%, then the better the value of discrimination.

IV. RESULTS AND DISCUSSION

A. Dataset

The data used in this study was taken from one of the online repositories, data world[12], this is the data collection from the website of the online lung cancer prediction system equipped with feedback from users. The site has been implemented during the period of August 2013, since then many people have visited the site.

The used dataset consists of 16 attributes and 309 rows of data. The attributes include 'GENDER', 'AGE', 'SMOKING', 'YELLOW_FINGERS', 'ANXIETY', 'PEER_PRESSURE', 'CHRONIC DISEASE', 'FATIGUE', 'ALLERGY', 'WHEEZING', 'ALCOHOL CONSUMING', 'COUGHING', 'SHORTNESS OF BREATH', 'SWALLOWING DIFFICULTY', 'CHEST PAIN', and 'LUNG_CANCER'. Furthermore, preprocessing data is done only by checking missing values and it turns out there is no missing value in the dataset.

The description of the number of tumor sufferers performed at this section aims to analyze the used data by observing the portion between the number of infected cancer and the number of uninfected cancers of overall number of datasets. The percentage composition of data can be seen in Fig. 2.

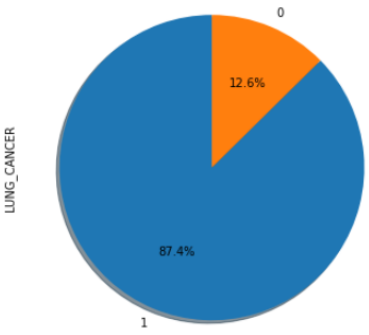

Fig. 2. Descriptive Data

Based on Figure 2, it can be known that the number of infected cancer (value = 1) is 87.4% of 309 data or 270 data. While the number of uninfected cancer (value = 0) is 12.6% of 309 data or 39 data. This indicates that the dataset is imbalanced.

There are many techniques to deal with imbalanced classification problems for imbalanced data, such as resampling the training data and developing other versions of the existing machine learning algorithm for the case's specific use. The simplest way to improve the accuracy for an imbalanced dataset in implementing classification techniques using the linear logistic model is to optimize the threshold parameter. The threshold parameter can be understood as a tolerant probability of a single data classified between two classes. Data preparation aims to specify independent variables and dependent variables. The next step is to divide the dataset into data test and data train; wherein this work, the data test size is set to be 30% of overall data.

B. Parameter Determination and Model Implementation

In this step, the process of elimination is done by eliminating variables whose significance value is more than 0.05. The process will continue until the model does not contain insignificant variables. The model used at this stage is logit because what will be done is the prediction of the probability of an event by matching the data on the logit function of the logistics curve. And this method is a common linear model used for binomial regression.

TABLE I
LOGIT REGRESSION RESULTS BEFORE ELIMINATION PROCESS

Logit Regression Results

Dep. Variable:	LUNG_CANCER	No. Observations:	216
Model:	Logit	Df Residuals:	201
Method:	MLE	Df Model:	14
Date:	Mon, 15 Feb 2021	Pseudo R-squ.:	0.5683
Time:	14:39:17	Log-Likelihood:	-35.133
converged:	True	LL-Null:	-81.382
Covariance Type:	nonrobust	LLR p-value:	1.278e-13

	coef	std err	z	P>|z|	[0.025	0.975]
GENDER	-0.8831	0.838	-1.053	0.292	-2.526	0.760
AGE	-0.0796	0.022	-3.543	0.000	-0.124	-0.036
SMOKING	1.2158	0.758	1.605	0.108	-0.269	2.701
YELLOW_FINGERS	0.7288	0.767	0.951	0.342	-0.774	2.232
ANXIETY	1.0657	0.887	1.202	0.229	-0.672	2.804
PEER_PRESSURE	1.0391	0.744	1.396	0.163	-0.420	2.498
CHRONIC DISEASE	1.9361	0.910	2.127	0.033	0.152	3.720
FATIGUE	2.4500	0.876	2.797	0.005	0.733	4.167
ALLERGY	1.5776	0.882	1.789	0.074	-0.151	3.306
WHEEZING	0.5778	0.975	0.593	0.553	-1.333	2.489
ALCOHOL CONSUMING	1.6698	0.974	1.715	0.086	-0.239	3.579
COUGHING	3.1252	1.240	2.520	0.012	0.695	5.556
SHORTNESS OF BREATH	-0.6321	0.866	-0.730	0.465	-2.330	1.065
SWALLOWING DIFFICULTY	2.9366	1.226	2.394	0.017	0.533	5.340
CHEST PAIN	0.5256	0.835	0.629	0.529	-1.111	2.162

As shown in Table.1, we focus on column P > |z| where the variables with significance value (p-value) are more than 0.05. The value of significance is the value of errors obtained during observation from statistical calculations [13]. After the elimination process, it can be seen in column P > |z| there is no variable with a significance value (p-value) greater than 0.05.

Next is to add constants to get a mathematical model. After that, the odds ratio is calculated to determine how much the risk factor is compared to lung cancer incidence. Or, in other words, how much influence a variable

has on the incidence of lung cancer. After that, determine the model to be used where the model used is logistic regression using a default threshold of 0.5. Then do the training data and predict the target value.

C. Result and Discussion

After implementing the explained technique above, it is obtained the following results as shown in Table.2.

TABLE II
LOGIT REGRESSION RESULTS AFTER THE PROCESS OF ELIMINATION AND ADDED CONSTANTS

Logit Regression Results

Dep. Variable:	LUNG_CANCER	No. Observations:	309
Model:	Logit	Df Residuals:	303
Method:	MLE	Df Model:	5
Date:	Mon, 15 Feb 2021	Pseudo R-squ.:	0.3514
Time:	14:39:17	Log-Likelihood:	-75.977
converged:	True	LL-Null:	-117.15
Covariance Type:	nonrobust	LLR p-value:	2.711e-16

	coef	std err	z	P>|z|	[0.025	0.975]
const	-3.2989	1.658	-1.990	0.047	-6.549	-0.049
AGE	0.0179	0.026	0.698	0.485	-0.032	0.068
CHRONIC DISEASE	1.9323	0.517	3.740	0.000	0.920	2.945
FATIGUE	1.7357	0.455	3.815	0.000	0.844	2.628
COUGHING	3.0482	0.562	5.421	0.000	1.946	4.150
SWALLOWING DIFFICULTY	3.4893	0.640	5.453	0.000	2.235	4.743

As shown in Table.2, the fitting data using the logit model with the optimization method to find the model's coefficients used Maximum Likelihood Estimation (MLE). MLE is a method of estimating parameters that maximize the likelihood function [14]. Next, Table 2 focuses on the 'coef' column where this column is the coefficients obtained for each independent variable so that the best model can be written with the following mathematical model.

logit(P) = -3.29 + 0.01 x Age + 1.93 x ChronicDisease + 1.73 x Fatigue + 3.04 x Coughing + 3.48 x Swallowing Difficulty

After obtaining mathematical equations, then calculating the odds ratio to determine how much influence a variable has on the incidence of people living with lung cancer, and received the following results as shown in Table.3.

TABLE III
ODDS RATIO AFTER THE ELIMINATION PROCESS AND ADDED CONSTANTS (LOGIT MODEL)

----------ODDS RATIO----------	
SWALLOWING DIFFICULTY	32.761528
COUGHING	21.077389
CHRONIC DISEASE	6.905130
FATIGUE	5.673011
AGE	1.018106
const	0.036922

As shown in Table.3, it can be seen that the factors that affect a person who has lung cancer are sorted from the most at risk, such as difficulty swallowing, coughing, chronic diseases, fatigue, and age.

After applying the model using a threshold of 0.5, which obtained an accuracy of 87.7%, then the ROC curve is determined in Figure 3 to identify how good the model is through statistical analysis.

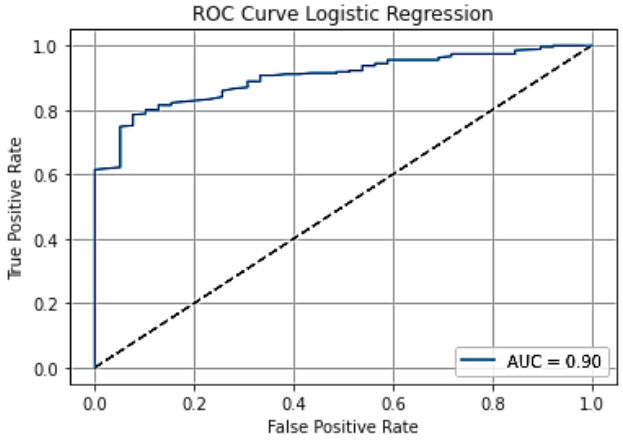

Fig. 3. ROC Curve – Logit Model

Based on Figure 3, the model obtained can be said to be a good model because in statistical analysis, the closer the ROC curve on the Y(0,1) line, the better the model. The ROC curve is a graphical representation of the relationship between sensitivity and 1-specificity. In medical research, the ROC curve is widely used to describe optimal diagnostic accuracy [15]. The AUC value obtained is 0.90 or 90%, whereas in statistical analysis, the AUC is usually used to get a discriminatory value. The closer the resulting value to 100%, the better the value of discrimination.

After applying the model, the next is to calculate each threshold's accuracy to know the comparison of accuracy values resulting from each threshold value shown in Figure 4.

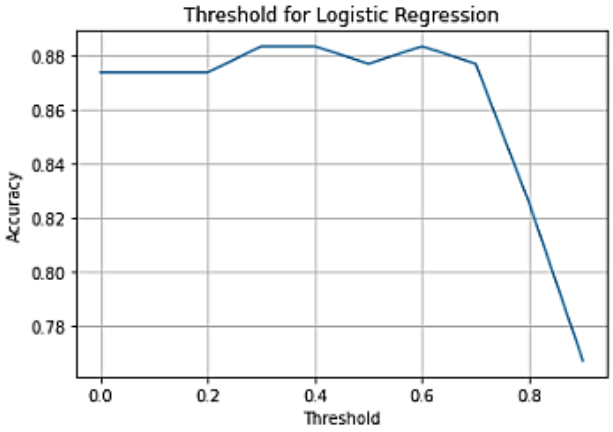

Fig. 4. Graph of Logit Threshold Values

Based on Figure 4, each threshold's accuracy is quite consistent before there is a significant decrease in accuracy value from the range of values 0.7-0.9. Threshold values with the top highest accuracy are obtained from the threshold 0.3, 0.4, and 0.6 with an accuracy value of 88.3%. The computation use threshold of 0.6 that is experimentally returning the highest accuracy, 88.3%.

In this experiment, two conditions were used: the first condition on parameter determination using the logit model and the second condition on parameter determination using GLM (Generalized Linear Model) model. Where in the second condition, this GLM model is used as a comparison. The results as shown in Table.4.

TABLE IV
GENERALIZED LINEAR MODEL REGRESSION RESULTS BEFORE THE ELIMINATION PROCESS

Generalized Linear Model Regression Results

Dep. Variable:	LUNG_CANCER	No. Observations:	216
Model:	GLM	Df Residuals:	201
Model Family:	Gaussian	Df Model:	14
Link Function:	identity	Scale:	0.070087
Method:	IRLS	Log-Likelihood:	-11.651
Date:	Mon, 15 Feb 2021	Deviance:	14.087
Time:	15:45:13	Pearson chi2:	14.1
No. Iterations:	3		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
GENDER	-0.0027	0.045	-0.060	0.952	-0.091	0.086
AGE	0.0025	0.001	2.316	0.021	0.000	0.005
SMOKING	0.0766	0.038	1.998	0.046	0.001	0.152
YELLOW_FINGERS	0.1389	0.048	2.923	0.003	0.046	0.232
ANXIETY	0.0877	0.053	1.667	0.095	-0.015	0.191
PEER_PRESSURE	0.0428	0.044	0.977	0.329	-0.043	0.129
CHRONIC DISEASE	0.0719	0.039	1.841	0.066	-0.005	0.149
FATIGUE	0.1484	0.046	3.222	0.001	0.058	0.239
ALLERGY	0.1707	0.040	4.223	0.000	0.091	0.250
WHEEZING	0.0411	0.044	0.939	0.348	-0.045	0.127
ALCOHOL CONSUMING	0.2051	0.050	4.142	0.000	0.108	0.302
COUGHING	0.1057	0.047	2.269	0.023	0.014	0.197
SHORTNESS OF BREATH	0.0490	0.046	1.067	0.286	-0.041	0.139
SWALLOWING DIFFICULTY	0.1323	0.048	2.755	0.006	0.038	0.226
CHEST PAIN	0.0253	0.041	0.613	0.540	-0.056	0.106

As shown in Table.4, we can know that the model used is GLM, and the method used is Iteratively Reweighted Least Square (IRLS). The IRLS method of weight is always recalculated on each optimization so that the overall analysis results are not affected by small data [16]. We focus on column P > |z| where we have to eliminate variables with a significance value (p-value) of more than 0.05. So the variables that need to be eliminated are gender, anxiety, peer pressure, chronic disease, wheezing, shortness of breath, and chest pain. The results of the Generalized Linear Model Regression Results after the elimination process and added constants as shown in Table.5.

TABLE V
GENERALIZED LINEAR MODEL REGRESSION RESULTS AFTER THE ELIMINATION PROCESS AND ADDED CONSTANTS

Generalized Linear Model Regression Results

Dep. Variable:	LUNG_CANCER	No. Observations:	309
Model:	GLM	Df Residuals:	300
Model Family:	Gaussian	Df Model:	8
Link Function:	identity	Scale:	0.072632
Method:	IRLS	Log-Likelihood:	-28.731
Date:	Mon, 15 Feb 2021	Deviance:	21.789
Time:	15:45:13	Pearson chi2:	21.8
No. Iterations:	3		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.1734	0.128	1.355	0.175	-0.077	0.424
AGE	0.0018	0.002	0.924	0.355	-0.002	0.005
SMOKING	0.0667	0.031	2.133	0.033	0.005	0.128
YELLOW_FINGERS	0.1682	0.035	4.745	0.000	0.099	0.238
FATIGUE	0.1764	0.035	5.084	0.000	0.108	0.244
ALLERGY	0.1640	0.033	4.930	0.000	0.099	0.229
ALCOHOL CONSUMING	0.1993	0.036	5.507	0.000	0.128	0.270
COUGHING	0.1025	0.034	3.032	0.002	0.036	0.169
SWALLOWING DIFFICULTY	0.1630	0.034	4.856	0.000	0.097	0.229

As shown in Table.5, it can be known that the factors that affect a person with lung cancer are age, smoking habits, yellow nails, fatigue, allergies, alcohol consumption, cough, and difficulty swallowing. When compared to Table 2, there are only 3 factors that are the same such as age, cough, and difficulty swallowing.

After applying the model using a threshold of 0.5 which obtains an accuracy of 0.90 or 90%, the ROC curve is obtained as follows that can be seen in Figure 5.

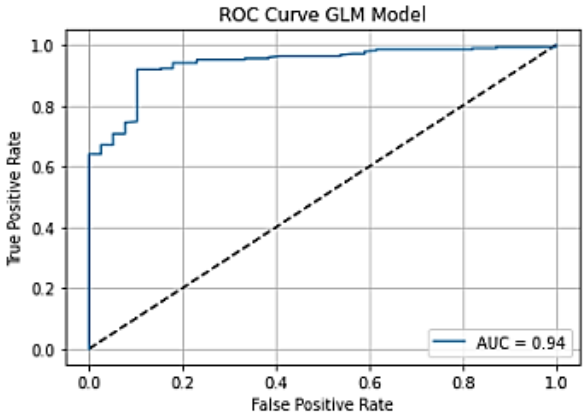

Fig. 5. ROC Curve – GLM Model

Based on Figure 5 it can be known that the model obtained can be said to be better than the previous model (Figure 6) because the AUC value obtained is 0.94 or 94% while the AUC value in the previous model is 0.90 or 90% where the resulting value is closer to 100% the better the value of discrimination. This indicates that the value of discrimination on the GLM model is better than the value of discrimination on the logit model.

Next is to try to calculate the accuracy of each threshold to know the comparison of accuracy values resulting from each threshold value. The results can be seen in Figure 6.

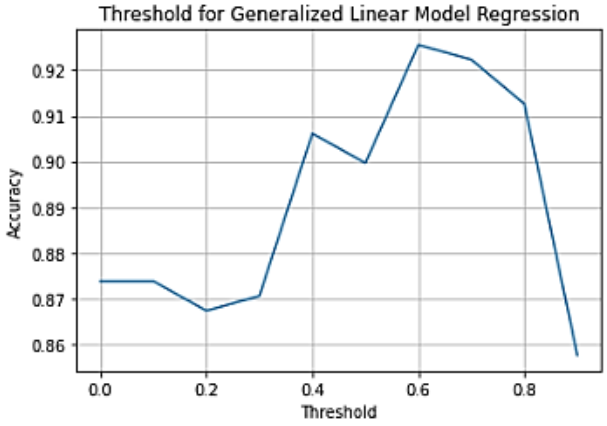

Fig. 6. Graph of GLM Threshold Values

Based on Figure 6, it can be known that the threshold value with the highest accuracy is 0.6 with an accuracy value of 92.5% The comparison of parameter determination between the two conditions used is the first condition using the logit model and the second condition using the GLM (Generalized Linear Model) model. The logit model obtained an accuracy value of 88.3% while on GLM models obtained an accuracy value of 92.5%, both are obtained from the threshold of 0.6. In the computation, the GLM increases the accuracy up to 4.2%, so the selection of models for parameter determination is also quite important.

V. Conclusion

Lung cancer is one of the most common causes of cancer death worldwide. The main reason behind the increasing of deaths from it is detecting the disease lately and faults in effective treatment. The detection and classification of lung cancer can help to enhance patient survival by detecting lung cancer at an earlier stage. Designing and analyzing machine learning techniques such as the logistic linear model for the classification of lung cancer without computing textural data or morphology. The proposed approach can predict lung cancer in its early stages which helps the survivability rate of the patients. The research concludes the variables or factors that significantly affect a person with lung cancer, such as difficulty swallowing, cough, chronic diseases, fatigue, and age. Logistic linear regression models for early prediction of lung cancer patients based on symptoms, habits, and history of health diseases obtained an accuracy of 0.877 or 87.7% and an AUC value of 90% by using a logit model on determining parameters and thresholds of 0.5 to see the likelihood that a person with a certain level of risk could have lung cancer. The selection of models for parameter determination and selecting thresholds used to get the best model is considered quite important because it affects the model's accuracy value. Hence, it is necessary to choose carefully.

References

[1] Kementrian Kesehatan RI. 2015. Situasi Penyakit Kanker. [Online] Available at:https://www.kemkes.go.id/resource/download/pusdatin/infodatin/infodatin-kanker.pdf
[2] Pusat Kajian Jaminan Sosial Universitas Indonesia (PKJS-UI). 2021. Kanker Paru, Kanker Paling Mematikan di Indonesia. [Online] Available at: https://pkjsui.org/wp-content/uploads/2021/05/PKJS-UI_Dialog-Pemangku-Kepentingan-Kanker-Paru.pdf
[3] Afrianto Yudi, Fauzy Muh. Farid, dan Setiawati Agustina. 2014. Kanker Paru. [Online] Available at: https://ccrc.farmasi.ugm.ac.id/?page_id=802
[4] Shen, S., Fan, Z., & Guo, Q. (2017). Design and application of tumor prediction model based on statistical method. Computer Assisted Surgery, 22(sup1), 232-239.

[5] Raghavendra, Patil G E., Sinchana, C G., Tejashwini, P., et al. 2020. Lung Cancer Prediction System Using Logistic Regression Approach. International Research Journal of Modernization in Engineering Technology and Science, 656-660.
[6] Hazra, A., Bera, N., & Mandal, A. (2017). Predicting lung cancer survivability using SVM and Logistic Regression Algorithms. International Journal of Computer Applications, 174(2), 19-24.
[7] Kasthuri, M., & Jency, M. R. (2020). Lung Cancer Prediction Using Machine Learning Algorithms on Big Data: Survey. International Journal of Computer Science and Mobile Computing, 9(10), 73-77.
[8] Redaksi Halodoc. 2018. Ketahui Perbedaan Tumor Jinak dan Tumor Ganas. [Online] Available at: https://www.halodoc.com/ketahui-perbedaan-tumor-jinak-dan-tumor-ganas
[9] Utomo, Setyo. 2009. Model Regresi Logistik untuk Menunjukkan Pengaruh Pendapatan per Kapita, Tingkat Pendidikan, dan Status Pekerjaan terhadap Status Gizi Masyarakat Kota Surakarta. Universitas Sebelas Maret Surakarta.
[10] Zaen, Nanida Jenahara. 2019. Diagnosis Penyakit Stroke dengan Metode Regresi Logistik Biner. UIN Sunan Ampel Surabaya.
[11] Rubiati, Meita Ariani. 2014. Penerapan Regresi Logistik Biner dan Analisis Dominan untuk Menganalisis Faktor-Faktor yang Berpengaruh terhadap Hipertensi. Institut Pertanian Bogor.
[12] Staceyinrobert. 2017. Survey Lung Cancer. [Online] Available at: https://data.world/sta427ceyin/survey-lung-cancer
[13] Yuniana, Deva Rizky. 2015. Perbedaan Nilai Alpha Dengan Nilai Signifikansi. [Online] Available at: http://fnistatistics.com/divisi_detail.php?id=114
[14] Nurlaila Dwi, Dadan Kusnandar, Evy Sulistianingsih. 2013. Perbandingan Metode Maximum Likelihood Estimation (MLE) dan Metode Bayes dalam Pendugaan Parameter Distribusi Eksponensial. [Online] Available at: https://core.ac.uk/download/pdf/326807809.pdf
[15] Kusmantoro Zaky Nur, Dr. Danardono, M.P.H., Ph.D. 2018. Akurasi Uji Diagnostik Menggunakan Luasan Bawah Kurva ROC Smoothed Empirical. [Online] Available at: http://etd.repository.ugm.ac.id/home/detail_pencarian/162253
[16] Setiawan, Fajar. 2012. Pemodelan Regresi Binomial Negatif dan Penerapannya. [Online] Available at: http://eprints.uny.ac.id/1413/1/PEMODELAN_REGRESI_BINOMIAL_NEGATIF_DAN_PENERAPANNYA.pdf