Prediction of Infecting Cancer-Based on Logistic Regression Model

Fusheng Luo [1, *]

[1]Department of mathematic school, University of Nottingham, Ningbo, China

Corresponding author: ssyfl2@nottingham.edu.cn

Keywords: Logistic Regression, Learning Vector Quantization, data normalisation, AUC-ROC curve

Abstract: Cancer is one of the most fatal contributors towards the increasing mortality rate of mankind. This represents an important topic to study for the sake of the welfare of humanity. However, this disease's traditional manual diagnosis and prognosis procedures are time-consuming, even for a professional medical practitioner. Thus, a model with robust power of predictions regarding the state of the tumour (i.e., probable cancer) would benefit most patients from the toxic side effects and additional medical services fees incurred by inessential treatment. To this end, the Logistic Regression Method is applied to derive a powerful model combining an algorithm from machine learning criteria – Learning Vector Quantization. There are two phases in building this model; phase 1 is the pretreatment of our data from Kaggle, including the process of normalization, classification, and feature selection. From feature selection, 14 variables are extracted based on their level of importance. Thereby, models are built on these 14 variables and one output Y, consisting of 0 or 1, derived from the classification process. These 14 variables have a significant impact towards the prediction process since they significantly reduce the work needed for the procedure. Phase 2 is applying the relevant methodology to produce our model and examine its efficiency. To test the ability of the trained logistic models to recognize cancer, we analyzed residual samples that were not previously used for the training procedure and correctly classified them in all cases. The evaluation of the model combines methods of the AUC-ROC curve as well as the confusion matrix, which are powerful statistical approaches. The AUC value after calculation is 0.9385144, which strengthens the validity and efficiency of the model. Besides, the confusion matrix reveals an accuracy of 0.9787 (out of 1). The repercussions of this model can be utilized to forecast the probability of cancer from concrete measurements of the tumour. This may refrain from the exorbitant expenditure on the usage of specific delicate medical machines, like X-rays. Moreover, this provides foundation statistics for the application of modern AI technology in the cancer prediction region.

1. Introduction

Cancer is a pronounced determinant of death around the world, accounting for almost 10 million deaths in 2020 [1]. Cancer's rising prominence as a leading cause of death partly reflects continuous declines in mortality rates of the population before the age of 70 years in 112 out of 185 countries worldwide [2]. The most common cancers are breast (around 4.7 million deaths), prostate (around 3 million deaths), lung (around 2.2 million deaths), and colorectum cancers. And there are various causes for this. The most publicly known determinants are tobacco use, alcohol consumption, and lack of physical activity, which contribute to one-third of deaths from cancer [3]. According to the articles published by the WHO (World Health Organization), evading risk factors and implementing existing evidential prevention strategies could benefit a lot. Indeed, between 30 and 50% of cancers can be prevented [3]. Besides, early diagnosis of lung cancer and suitable treatment can effectively improve the survival rate of patients by 20% [4]. At present, the

survival condition of lung cancer patients is not optimistic, though diverse methods of treatments are being applied to clinical practice [5]. However, within those methods, Artificial Intelligence (AI) has been shown as an effective weapon in cancer diagnosis [6]. There is a long history of AI being applied in this field. Recently, in 2019, Google coined its lung cancer detection AI, which has a strong efficiency matched up with six human radiologists [7]. Therefore, an algorithm for the distinguishment of tumour becomes quite compulsory for the analysis procedure. Accordingly, the model we built in this paper to help detect the type of tumour from benign to malignant is of high importance on account of its effect on bringing high probabilities for fruitful treatment. Based on several sorts of data about material features of the tumour, the accurate prediction of a specific type of tumour (i.e., Malignant (M) or Benign (B)) has an immense impact on the consequent medical procedures dealing and the higher possibility for survival of patients.

2. Data Description and exploratory analysis

This paper utilizes data collected from Kaggle, a site that provides authorized datasets to its users. The data under examination pertains to a specific type of tumour, defined as a mass or cluster of abnormal cells within the body. Tumours can be classified as either malignant (M) or benign (B), and only malignant tumours typically result in cancer. This paper focuses on malignant tumours. The dataset contains 10 tumour descriptors, including radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension. Each feature is presented in three categories: mean value, worst value, and standard error. Of the 10 variables, radius, perimeter, and area analyze tumour size. It is thought that the likelihood of a benign tumour increases as the size of the tumour expands. Additionally, benign tumours generally exhibit smoother features and smoother edges. Before the biological model is constructed, the first thing to do is give a brief picture of the relationship between the variables and find any apparent relationship between them. To begin with, we want to know the distribution of the diagnosis of our data.

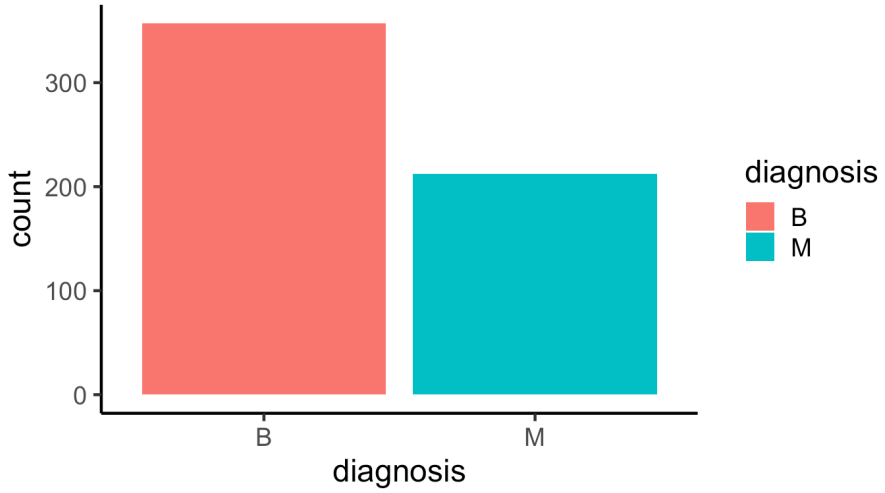

Figure 1: Bar chart of the distributions of the diagnosis

As Figure 1 shows, among 569 groups of data, more than 300 groups give the benign diagnosis, and nearly 200 groups present the malignant diagnosis. From this perspective, the numbers are distributed roughly evenly, and this makes our model powerful enough because of the comprehensiveness of the information contained in the density of datasets.

What's more, it's our interest to dig into the connection between all these independent variables to better comprehend this dataset. Thus, a picture of the correlation has been plotted below:

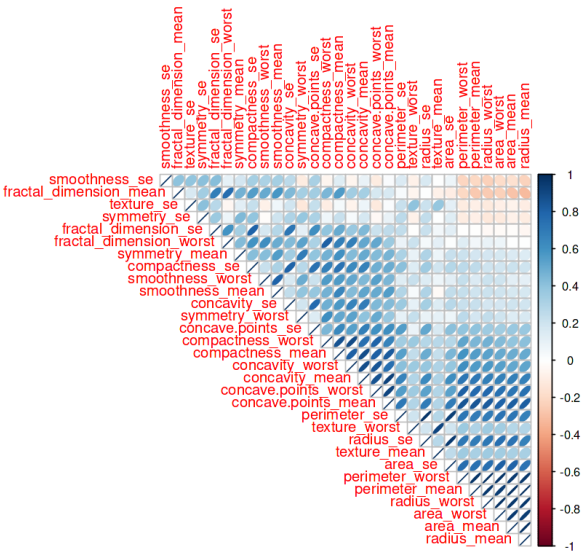

Figure 2: Correlation plot between different variables

Figure 2 displays the correlation among the independent variables, with the colour bar on the right indicating a range from -1 to 1. A linear negative correlation is represented by -1, while a linear positive correlation is represented by +1. The graph's colours indicate the degrees of connection, with the darkest blue representing +1 and the darkest red representing -1. The transitions from red to blue indicate different levels of relationships. We have excluded comparisons between variables and themselves since it would always be 1, which is denoted by forward slashes. Upon examining the graph, we observe that most of the correlations among the 30 variables (consisting of 10 variables with 3 different data standards each) are positive. For instance, the navy-blue colour represents the relationship between the three variables of compactness and the three fractal dimensions. This relationship implies that there are strong connections between a tumour's compactness and its complexity in dimensions (degree of shapelessness). It is worth noting that the relationship between smoothness and perimeters or areas is negative. This means that a highly smooth tumour is less likely to be significant in size.

3. Methodology

Based on the processed data, two main machine learning methods are applied to construct an appropriate mathematical model that can reasonably be used to predict the type of tumour (the probability of cancer) with all the material information. Those methods are Learning Vector Quantization and Logistic regression.

3.1 The Learning Vector Quantization (LVQ)

Artificial intelligence approaches, particularly machine learning and deep learning are increasingly reconstructing the full spectrum of clinical management for cancer (gastric cancer especially) [8]. Learning vector quantization [9] is one kind of artificial neural network algorithm that was originally designed to be applied in the fields of biological models of neural systems. LVQ is based on a prototype supervised learning (given both labels and data) classification algorithm and can deal with the multiclass classification problem. In general, the architecture of LVQ has two layers: The input layer and the Output layer; one illustration is shown in Figure 3 below:

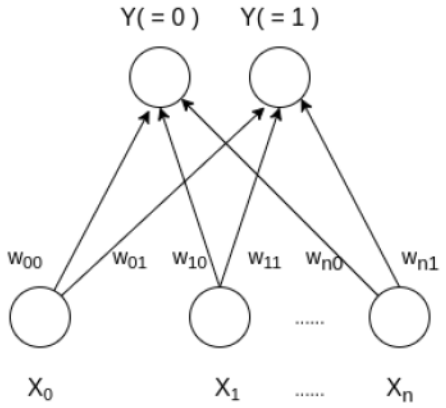

Figure 3: schematical representation of the LVQ algorithm

Given the initial input data, LVQ, as a type of clustering algorithm, divides inputs into numbers of groups (here Y=0 and Y=1). These output values are also known as "node", which is the representation of a certain type of group of inputs. Thus, whenever the new input comes in, it can be arranged according to the type of group that qualifies for the input value.

3.2 Logistic Regression

A type of statistical model (*logit model*) which is quite beneficial for classification problems and predictive analytics and has been successfully applied to cancer classification problems [10]. A fundamental concept of logistic regression is the odds ratio (OR), which represents the odds that an outcome will occur given a particular event, compared with the odds of the same outcome happening in the absence of that event. Equivalently speaking, it's the probability of success divided by the probability of failure. If the OR is greater than 1, the event is associated with a high probability of producing a particular result, and vice versa. Besides, we want to derive a probability of the outcome, which is deemed to be bounded between 0 and 1. Therefore, the logistic function [11] was developed as follows:

$$\text{logit}(\pi) = \frac{1}{1+e^{-\pi}} \tag{1}$$

$$\ln(\frac{\pi}{1-\pi}) = \beta_0 + \beta_1 \times x_1 + \cdots + \beta_k \times x_k \tag{2}$$

In this equation, $\text{logit}(\pi)$ is the dependent variable, and x is the independent variable. Moreover, maximum likelihood estimation (MLE) predicts the beta variables here. This logistic method tests different beta values through many iterations to optimize for the best fit of log odds. Once all the coefficients are found, the conditional probabilities for each observation can be calculated, logged, and summed together to produce an estimated probability. For binary classification (the same as we are doing in this paper, it's just a method of dividing the output into the binary form, i.e., 0 and 1), a calculated probability less than 0.5 would give a result of 0, and one greater than 0.5 would

yield a 1. In this way, Logistic Regression benefits a lot in this classification problem, and based on this, we can build a robust model. What's more, there are paths to examine the robustness and efficiency of our model; the method applied here is named the *AUC-ROC test*.

4. Results

4.1 Data Pretreatment

Extracting correlated biological information from massive datasets is a pronounced challenge in modelling biological research. Different aspects of data may hamper their interpretations in the research [12]. From this stance, data pretreatment methods become incredibly important because they can correct large but prolix datasets by emphasizing the biological information within and, in this way, increasing their biological interpretability.

There are various methods of data pretreatment; the mainstream methodologies contain centring, autoscaling, P*areto scaling, range scaling, vast scaling, log transformation, and power transformation*. All the methods have been tested in Van den Berg's paper, and more content can be found regarding the advantages and drawbacks of different methods. At all events, the method applied in this paper is *min-max normalization*, which compares metabolites relative to the biological response range. In this way, all metabolites become equally crucial despite the fact that inflation of the measurement errors may occur, and the model is quite sensitive to outliers.

4.2 Normalization

Since the *min-max normalisation* [13] is applied, it follows the formula downwards:

$$\widetilde{x_i} = \frac{x_i - x_{i_{min}}}{x_{i_{max}} - x_{i_{min}}} \tag{3}$$

Here, i is the label of order, in the whole sequence of X. And $\tilde{x}$ is denoted as x after the transformation. This helps to transfer all the data into scales between 0 and 1, and based on this, subsequent calculation and modelling can be done with all comparable data since we focus more on the relative value of data rather than the original one. What's more, we also classify malignant tumours as "1", and "0" for benign ones.

4.3 Feature Selection

In this paper, the LVQ method is applied to determine the importance of the variables and rank them accordingly. From this perspective, the graph below gives out the 20 most important variables out of 30.

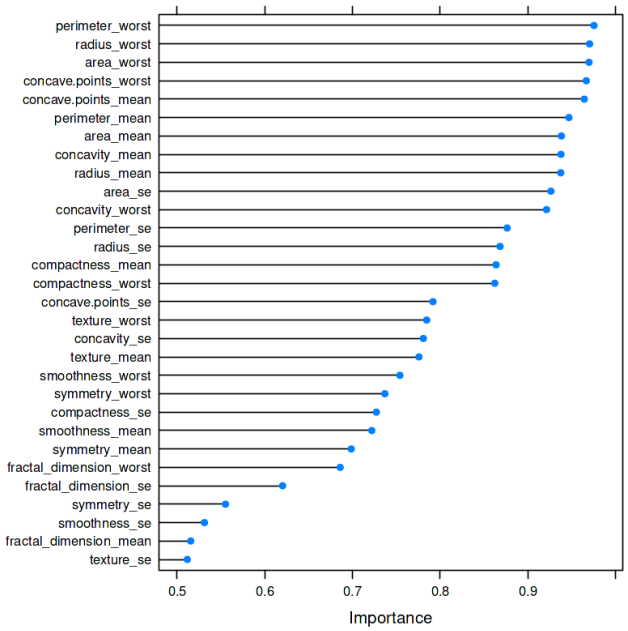

Figure 4: The importance rankings of variables in the graph form

Figure 4 above is also a graphic version of the importance of all the 30 variables; among those, the 20 most important variables are shown with the value of importance from 0 to 1 in descending order. Besides from the graph, we can see that almost 14 variables (perimeter_worst, radius_worst, area_worst, concave.points_worst, concave.points_mean, perimeter_mean, area_mean, concavity_mean, radius_mean, area_se, concavity_worst) have the importance over 0.8, which satisfy what we need with regard to the number of variables in the model. Hence, we set a threshold of 0.8 to distinguish the degree of importance of all 30 variables. From this perspective, we obtain the 14 variables which have shown robust importance to our model. These variables also have massive contributions towards the modern medical area. Select the most critical variables and the process of data acquisition can be reduced mainly since 14 variables are now in need rather than 30. This can also increasingly cut the expenses needed for the procedure.

4.4 The Validation Set Approach

To make our model robust enough, the validation set approach is applied here to divide all data into two parts: the training data and the test data. We use a split ratio of 0.8, which corresponds to the importance threshold, for uniformization. In this way, the data is randomly split into 10 parts, and 8 parts of them are transferred into groups of training data.

4.5 Model Building

Based on their importance level, we have found the variables we need to build the desired model. Logistic Regression is an excellent choice here. The picture below is the feedback information regarding the Logistic Regression model we're building in the R program.

Table 1: Output of R in terms of the summary of the Logistic Regression Model

Coefficients:			
(Intercept)	radius_mean	perimeter_mean	area_mean
-8.342	-547.821	466.342	101.270

compactness_mean	concavity_mean	concave.points_mean	radius_se
-40.994	-22.874	19.018	17.870
perimeter_se	area_se	radius_worst	perimeter_worst
-74.591	189.806	105.106	-26.386
area_worst	compactness_worst	concavity_worst	concave.points_worst
-75.142	5.453	27.840	11.659

Table 1 tells a lot regarding the Logistic Regression Model; for instance, the degree of freedom is 425, Null Deviance is 561.8, and Residual Deviance is 60.76. From this, we can calculate the coefficient of determination $R^2 = \frac{561.8-60.76}{561.8} = 0.8918$. The *AIC* value only matters in comparison. Besides, this gives the figure of all the coefficients in the logistic model, namely, the weights w. With 425 degrees of freedom, we then construct our model as in (2):

$$\ln\left(\frac{y}{1-y}\right) = (-547.8) \times x_1 + (466.3) \times x_2 + (101.2) \times x_3 + (-41.0) \times x_4 + (-22.9) \times x_5 + (19.0) \times x_6 + (17.9) \times x_7 + (-74.6) \times x_8 + (189.8) \times x_9 + (105.1) \times x_{10} + (-26.4) \times x_{11} + (-75.1) \times x_{12} + (5.5) \times x_{13} + (27.8) \times x_{14} + (11.7) \times x_{15} \quad (4)$$

And from (1), $y = 1$ if $logit(y) > 0$ and $y = 0$ if $logit(y) < 0$. And, x_i denotes the ith input from the data, and from $i = 1$ to $i = 14$, each denotes "*radius_mean, perimeter_mean, area_mean, compactness_mean, concavity_mean, concave.points_mean, radius_se, perimeter_se, area_se, radius_worst, perimeter_worst, area_worst, compactness_worst, concavity_worst, concave.points_worst"* respectively.

4.6 Model Evaluation

Evaluation of this model is equally important as naming the letter one sends out. First, based on the training data, we must utilize the test data to examine the efficacy of our model.

Table 2: Predictions from the model in terms of different patients' data

4	10	13	15	20	26	29
0.9309070	0.9999992	0.9999988	0.9772015	0.1164363	1.000000	0.9999994

Table 2 consists of the predictive results of our modelling applied to test data. Remember that there are 141 samples (patients) in total. This gives the probability of one patient having a malignant tumour in the forecast; for example, sample number 4 has a probability of 0.93 of getting cancer, and patient 25 would have cancer, whereas patient 20 has a relatively low tendency to get cancer. However, given the sample size of our model is not massive enough, more needs to be considered when evaluating this model. To take a step further, we use the concept of a confusion matrix to improve our test.

A confusion matrix [14] is a situation analysis table that summarizes the classification model's prediction results in machine learning and summarizes the records in the data set in the form of a matrix according to the natural category and category judgment predicted by the classification model. The rows of the matrix represent the actual values, and the columns of the matrix represent the predicted values.

Table 3: confusion matrix

		Prediction value	
		0	1
True value	0	87	1
	1	2	57

From table 3, one can spot that there exist 128 (51 + 87) accurate predictions, and this reveals an accuracy of around 0.9787. Note that the accuracy can be calculated using the below formula,

$$acc = \frac{True}{True+False} \tag{5}$$

Where True is the number of the uniformity between predictions and absolute values.

ROC-AUC [15] is another classic graphic examination of the evaluation of one model. ROC represents receiver-operating characteristic curves, and AUC is short for "area under the curve," which measures the size of the area circumscribed by the ROC curve. By convention, the x-axis of the ROC plot is the false positive rate (FPR), and the y-axis is the true positive rate (TPR); both are derived from the formula.

$$speciality = \frac{TN}{TN + FP} \tag{6}$$

$$FPR = \frac{FP}{FP + TN} = 1 - speciality \tag{7}$$

$$TPR = \frac{TP}{FP + TN} \tag{8}$$

Where FP is the false positive, it is supposed to be false though the model gives an accurate prediction. TN is the actual negative, the model gives an excellent prediction of a negative sample.

The AUC-ROC curve is a performance measure for classification problems under various threshold settings. The ROC is the probability curve, and the AUC represents the degree or measure of separability, which tells us how many models can distinguish between classes. The higher the AUC, the better the model is at predicting 0 to 0 and 1 to 1. In fact, the higher the AUC, the better the model is at distinguishing between patients with and without disease (cancers).

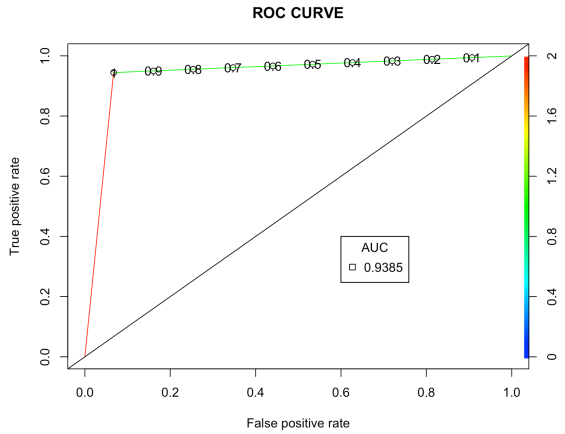

Figure 5: The ROC curve (True Positive rate v.s. False Positive rate)

According to Figure 5, the program gives out a result of 0.9385144 as the AUC value, which increasingly supports the validity of our modelling [15].

5. Conclusions

The early detection and appropriate treatment of cancer are critical for successful rehabilitation. Unfortunately, identifying the early stages of a tumour can be a challenging and expensive task. To address this issue, we've utilized machine learning to predict cancer using a basic model. We've built a Logistic Regression model based on pre-measured data to predict whether a tumour is benign or malignant. This model includes variables such as detailed measurements of the tumour's radius, perimeter, area, compactness, concavity, and concave points, along with statistical properties such as mean, variance, maximum, and minimum value. By selecting the most important variables, we were able to significantly reduce the time and money required for early prediction. Without feature selection, the accuracy of our model would be greatly reduced. Our model has an accuracy of 0.9787 and an AUC value of 0.9385, statistically validating its effectiveness. As long as the necessary data is available, the model can be used in various cases and can contribute to the development of basic AI technology in the field of cancer prediction. However, it requires complete data with various information about the tumour, which can be added and accessed for more detailed assessments.

Reference

[1] Ferlay J, Ervik M, Lam F, et al. Global Cancer Observatory: Cancer Today. Lyon: International Agency for Research on Cancer. IARC; 2018[J]. 2020.

[2] Cancer[EB/OL]. World Health Organization, World Health Organization, 2022-02-03. (2022-02-03)[2023-11-11].

[3] Sung H, Ferlay J, Siegel R L, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries[J]. CA: a cancer journal for clinicians, 2021, 71(3): 209-249.

[4] Lin H T, Liu F C, Wu C Y, et al. Epidemiology and survival outcomes of lung cancer: a population-based study[J]. BioMed Research International, 2019, 2019.

[5] Kuzniar T J, Masters G A, Ray D W. Screening for lung cancer-a review[J]. Medical Science Monitor, 2004, 10(2): RA21-RA. 30.

[6] Huang S, Yang J, Shen N, et al. Artificial intelligence in lung cancer diagnosis and prognosis: Current application and future perspective[C]//Seminars in Cancer Biology. Academic Press, 2023.

[7] Svoboda E. Artificial intelligence is improving the detection of lung cancer[J]. Nature, 2020, 587(7834): S20-S20.

[8] Wang Z, Liu Y, Niu X. Application of artificial intelligence for improving early detection and prediction of therapeutic outcomes for gastric cancer in the era of precision oncology[C]//Seminars in Cancer Biology. Academic Press, 2023.

[9] Learning vector quantization[EB/OL]. GeeksforGeeks, 2023-01-07. (2023-01-07)[2023-11-10].

[10] Zhou X, Liu K Y, Wong S T C. Cancer classification and prediction using logistic regression with Bayesian gene selection[J]. Journal of Biomedical Informatics, 2004, 37(4): 249-259.

[11] Logistic regression in machine learning[EB/OL]. GeeksforGeeks, 2023-07-14. (2023-07-14)[2023-11-10].

[12] Van den Berg R A, Hoefsloot H C J, Westerhuis J A, et al. Centering, scaling, and transformations: improving the biological information content of metabolomics data[J]. BMC genomics, 2006, 7: 1-15.

[13] Data normalization in Data Mining[EB/OL]. GeeksforGeeks, 2023-02-02. (2023-02-02)[2023-11-10].
[14] Confusion matrix in machine learning[EB/OL]. GeeksforGeeks, 2023-03-21. (2023-03-21)[2023-11-10].
[15] AGARWAL R, Roc Curves & AUC: The Ultimate Guide[EB/OL]. Built In, 2022-08-18. (2022-08-18)[2023-11-10].