JOURNAL OF TECHNOLOGY MANAGEMENT AND BUSINESS
ISSN: 2289-7224 e-ISSN: 2600-7967

JTMB

Vol. 11 No. 2 (2024) 72-82
https://publisher.uthm.edu.my/ojs/index.php/jtmb

The Uses of Multiple Linear Regression as A Predictive Model for Factors of Breast Cancer in Malaysia

Muhammad Ammar Shafi[1]*, Zulfana Lidinillah[1], Shawal Iskandar Sabri[1], Nor Faezan Abdul Rashid[2], Mohd Saifullah Rusiman[3], Nur Ain Ebas[4]

[1] *Department of Technology and Management, Faculty of Technology Management and Business, Universiti Tun Hussein Onn Malaysia, 86400, Batu Pahat, Johor, MALAYSIA*
[2] *Surgery department, Hospital Al-Sultan Abdullah, Universiti Teknologi MARA, 42300 Bandar Puncak Alam, Selangor, MALAYSIA*
[3] *Department Mathematics and Statistics, Faculty of Applied Sciences and Technology, Universiti Tun Hussein Onn Malaysia Edu Hub Pagoh, 84600 Muar, Johor, MALAYSIA*
[4] *Department of Civil Engineering, Faculty of Civil Engineering and the Built Environment, Universiti Tun Hussein Onn Malaysia, 86400, Batu Pahat, Johor, MALAYSIA*

*Corresponding Author: zulfanalidinillah99@gmail.com, ammar@uthm.edu.my
DOI: https://doi.org/10.30880/jtmb.2024.11.02.006

Article Info

Received: 24 July 2024
Accepted: 7 September 2024
Available online: 10 December 2024

Keywords

Breast cancer, predicting model, multiple linear regression, mean square error

Abstract

The application of multiple linear regression analysis has significantly increased in popularity among researchers, becoming a predominant model for the analysis of data associated with complex phenomena. This study specifically focuses on the prediction of breast cancer factors using linear regression. Data was collected from 569 breast cancer patients who received treatment at the general hospital in Malaysia, with the secondary data being recorded by nurses and doctors. To ascertain the factors influencing breast cancer, research was conducted at the general hospital in Malaysia, which examined four independent variables, each with various combinations of variable types. The primary aim of this study was to identify the factors that significantly influence breast cancer factors at general hospital. All collected data will be analyzed using the Statistical Package for Social Science (SPSS) software. The analysis will include tests for data normality and the calculation of coefficients to meet the study's objectives. The findings indicate that breast cancer severity is significantly influenced by radius, as determined through multiple linear regression analysis. The conclusion of this chapter will present a comprehensive summary of the research findings, as well as acknowledge the limitations of the study. Furthermore, recommendations for enhancing other predicting modeling techniques in future breast cancer research are included.

1. Introduction

Regression analysis has emerged as a pivotal model for data analysis, garnering its popularity from a multitude of factors. This statistical technique generates a mathematical equation, delineating the relationship between the dependent and independent variables. Its explanatory capacity is notably high due to its multidimensional nature. It is readily available within computer software packages and presents a comprehensible approach. Its application extends across a broad spectrum, including applied sciences, economics, engineering, computer science, social sciences, and various other disciplines (Agresti et al., 1996).

Linear regression represents a statistical methodology designed to encapsulate and investigate the relationships between two continuous (quantitative) variables. Specifically, one variable, referred to as *Y*, is considered the dependent variable, response, or outcome. Concurrently, another variable, denoted *X*, is identified as the independent variable, predictor, or explanatory variable. The underlying principle of linear regression posits that the expected value of the dependent variable can be approximated by a linear function of the independent variables (Draper, N.R., & Smith et al., 1998).

Particularly for applications in the social and biological sciences, the usage of specialized statistical methods for categorical data has grown significantly in recent years (Agresti, 1996).

One of the most used analysis methods in market research is regression analysis. It enables the analysis of dependent and independent variable interactions by researchers. The outcome is the dependent variable, and the independent factors are what lead to those results (Sarstedt & Mooi, 2019). The applied sciences, economics, engineering, computer science, social sciences, and other sectors have all made extensive use of it (Agresti, 1996).

For all statistical techniques, statistical regression analysis is the most frequently employed and has a wide range of applications to many real-world issues (Junaid,2022). One statistical method for examining correlations between variables is regression analysis (Sykes, 1993).

Many models, including multiple regression, quadratic regression, cubic regression, logit model, probit model, exponential model, growth model, neural network regression, and fuzzy regression, are now available as a result of regression analysis. The medical field can benefit from the application of statistical tools (Shafi & Rusiman, 2015). Regression analysis can be used to estimate the relative strength of the effects of various independent variables on a dependent variable, determine whether one independent variable or a set of independent variables has a significant relationship with the dependent variable, and make predictions (Sarstedt & Mooi, 2019).

2. Literature Review

In the year 2018, a global tally of 9.6 million cancer fatalities was documented, alongside an anticipated 18.1 million new cancer incidences (World Health Organization, 2018). The International Agency for Research on Cancer (IARC) has delineated breast cancer as the predominant type of malignancy across all ethnic groups among women, resulting in 627,000 fatalities in the same year (World Health Organization). The organization further projected 2.1 million new diagnoses of breast cancer in 2018, equating to nearly one in every four cancer cases diagnosed in women (Bray et al., 2018). Despite the lower prevalence seen in Asian countries, it was noted that one in every twenty women in Malaysia faces a risk of breast cancer over their lifespan. Astonishingly, a significant percentage, 75%, of breast cancer cases exhibited no identifiable risk factors (Lee et al., 2019).

A recent study by Xu et al. (2023) has highlighted breast cancer as the foremost malignancy among women in Malaysia, representing 34.1% of all new cancer cases between 2011 and 2016. In Malaysia, the risk of breast cancer for a woman stands at approximately one in twenty, with an annual incidence of 8,418 new cases and 23 new cases each day. The projected number of new cancer cases in Malaysia is anticipated to more than double by the year 2040, reaching 48,639 in 2021 (IARC, 2021).

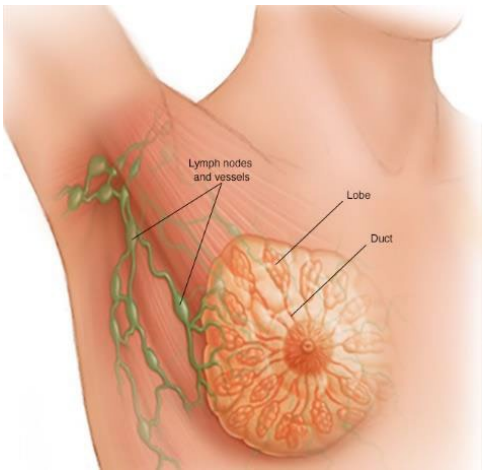

Fig. 1 *Breast cancer*

J. of Technology Management and Business Vol. 11 No. 2 (2024) p. 72-82

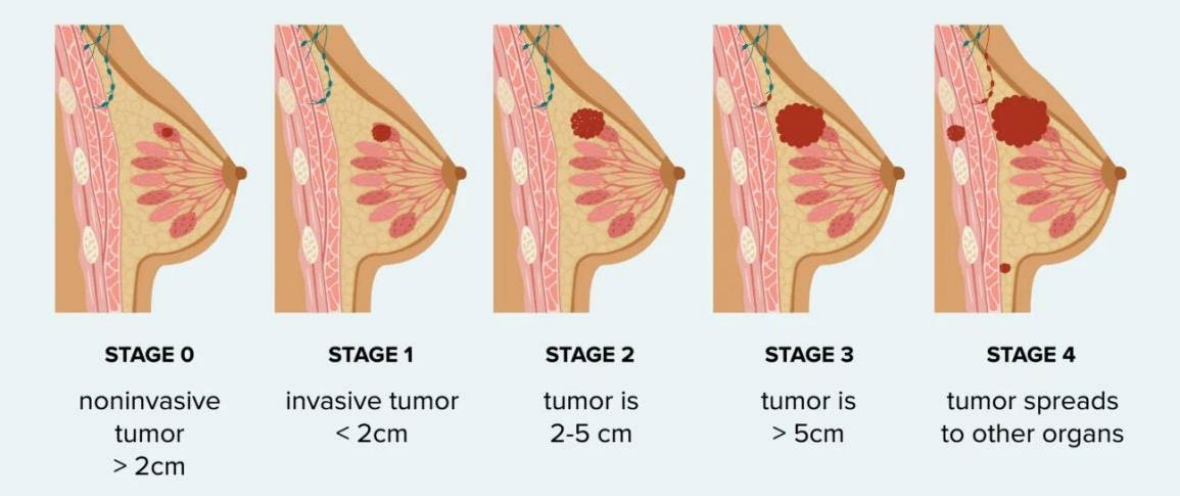

Fig. 2 *Stages of breast cancer*

The cells lining the ducts are where most breast malignancies start (ductal tumors). A tiny percentage of lobular malignancies originate in other tissues, whereas the majority start in the cells lining the lobules. Breast cancer can be classified as either invasive or non-invasive. Non-Invasive Breast Cancer Cells that stay inside the ducts and don't spread to the breast's surrounding connective and fatty tissues. Among non-invasive breast cancers, ductal carcinoma in situ (DCIS) is the most prevalent type. According to Sharma et al. (2010), invasive breast cancer cells penetrate the duct and lobular wall to spread across the breast's surrounding fatty and connective tissues.

Breast cancer symptoms or risk factors can identify the disease. The six factors of breast cancer are age, radius, area, smoothness, compactness, and concavity. Breast cancer is mostly caused by inherited genetic abnormalities in the breast cancer susceptibility genes BRCA1 and BRCA2, as well as a personal or family history of the disease (Bogdanova et al., 2013). Obesity, hormone therapy use, increased breast tissue density, alcohol consumption, and sedentary lifestyles are potential additional risk factors for breast cancer (Majeed et al., 2014).

Based on the extent of the disease, a stage is assigned to cancer at the time of diagnosis. The stage aids medical professionals in determining the best course of action and prognosis. Breast cancer stages can be broadly classified as invasive or in situ (not invasive). Stages can be given numerical designations (0 to IV) and detailed descriptions. Stage 0 is the non-invasive stage of a tumor, meaning that there is no indication of the tumor's invasion of the surrounding tissues and that both cancerous and non-cancerous cells are contained within the borders of the breast portion where the tumor is growing (Bednarek et al., 1997).

Stage I pertains to invasive breast cancer; wherein microscopic invasion may occur. There are two categories: 1A stage and 1B stage. Stage 1B represents a tiny collection of cancer cells larger than 0.2 mm detected in a lymph node, whereas category 1A describes a tumor up to 2 cm in size that does not involve any lymph nodes (Segal et al., 2001). There are two more categories in stage II: 2A and 2B. Stage 2A reports that while there is no malignancy in the breast, the tumor is discovered in the sentinel or axillary lymph nodes. The tumor's size might range from less than 2 cm to more than 5 cm. On the other hand, stage 2B indicates that although the tumor may not be able to reach the axillary lymph nodes, it may be larger than 5 cm (Moran et al., 2014). Similarly, stage III has been further subdivided into 3A, 3B, and 3C subcategories. Whereas stage 3B indicates that a tumor of any size may have caused swelling or an ulcer on the skin of the breast and may have spread to up to nine axillary lymph nodes or to sentinel lymph nodes, stage 3A describes that no tumor is found in the breast but that it may be found in four to nine axillary lymph nodes or in sentinel lymph nodes. Inflammatory breast cancer, or stage 3B, is characterized by red, heated, and swollen breast skin. However, according to Jacquillat et al. (1990), stage 3C denotes the tumor's progression to ten or more axillary lymph nodes, as well as the lymph nodes above and below the collarbone. The advanced and metastatic stage IV of cancer refers to the disease's spread to other body organs, including the brain, liver, lungs, and bones (Neuman et al., 2010).

While many men and women with breast cancer do not have symptoms, breast cancer is occasionally discovered after symptoms start to show. This explains the significance of routine breast cancer screening. Exams and tests used to identify a disease in individuals without any symptoms are referred to as screening. Finding breast cancer early, before symptoms appear (such as a palpable lump in the breast), is the aim of screening testing. Various screening tests are available, including digital mammography, thermography, ultrasound, magnetic resonance imaging (MRI), and screening MRI (Budh & Sapra, 2022). Women with BRCA gene mutations,

a strong family history of breast cancer, and a history of chest radiation therapy should undergo annual mammography and MRIs, as well as perhaps 6-monthly scans (Ma et al., 2019).

3. Methodology

A regression model with simulated data will be used where the data are generated using Monte Carlo simulation [19]. For the real data of breast cancer as secondary data, the data were obtained from the general hospital in Malaysia. It involves around 569 patients as respondents for breast cancer, and the data were collected and recorded by doctors and nurses using cluster sampling. As continuous data, the dependent variable is tumour size, and six factors of breast cancer are the independent variables. The software used to obtain accurate results was the Statistical Package for Social Sciences (SPSS).

3.1 Descriptive Statistics

Descriptive statistics serve the purpose of organizing data by establishing the relationships between variables within a sample or population (Kaur, 2018). Descriptive analysis is segmented into two primary categories: measures of central tendency and measures of variability. In the context of this study, a measure of central tendency was employed to determine the mean.

The interpretation of Wiersma's mean for agreeableness is presented in Table 1. An average mean value of 1.00 to 2.33 is classified as weak, values ranging from 2.34 to 3.67 are categorized as moderate, and values falling between 3.68 and 4.33 are high.

Table 1 *Agreeableness level according to mean interpretation by Wiersma Samsudin, Awang & Ahmand (2017)*

Mean	Central Tendency Level
1.00 – 2.33	Weak
2.34 – 3.67	Moderate
3.68 – 5.00	High

3.2 Multiple Linear Regression

Sir Francis Galton, a prominent figure of the nineteenth century, pioneered the development of regression analysis. Through his research on the correlation between the heights of parents and their children, he observed a phenomenon known as “reversion to the mean”. This trend suggested that the heights of children born to parents of varying stature tended to return to an average height, often perceived as a level of “mediocrity”. To elucidate this trend, Galton devised a mathematical framework.

In the realm of statistics, regression analysis refers to the exploration of the statistical relationships between variables (Kutner, 2004). It is important to note that many statistical tests predicate their results on certain assumptions regarding the variables under study. For the regression model to be executed, it is imperative that these assumptions are satisfied (Montgomery and Peck, 1992). The list of assumptions for the multiple linear regression model, proposed by Brant (2007), includes:

i. The dependent variable, denoted by y, exhibits a linear and additive pattern with respect to its mean in each stratum defined by x.
ii. It is postulated that the observations of y are statistically independent.
iii. The variance of y within each stratum is consistent for all values of x.
iv. The distribution of y across the x-strata is presumed to be normally distributed.

For the purpose of conducting analyses utilizing multiple linear regression, it is imperative that the data utilized adheres to specific assumptions. In the context of this study, three particular assumptions were scrutinized: constant variance, normality, and multicollinearity. It is crucial for these assumptions to be satisfied to guarantee the reliability of the results obtained. Numerous articles have elucidated that their conclusions are derived from the fulfillment of these assumptions in the statistical tests.

Multiple linear regression encompasses the incorporation of multiple predictor variables, which can range from models consisting of only two predictors (first-order models) to models with an arbitrary number of predictors. The mathematical formulation of a multiple linear regression model is as follows:

J. of Technology Management and Business Vol. 11 No. 2 (2024) p. 72-82

$$\overline{Y} = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_j x_j \quad (1)$$

Where:
$\beta_o, \beta_1, \ldots, \beta_j$ are constants
$X_{i1}, \ldots, X_{i,j}$ are unknown parameter/ independent variables
i = 1,..., n

In the context of statistical analysis of variance, the sums of squares include the Sums of Regression (SSR), Sums of Error (SSE), and Sums of Total (SST). Sums of Squares for the Analysis of Variance:

$$SSR = bX'Y - \left(\frac{1}{n}\right)Y'JY \quad (2)$$

$$SSE = (Y - Xb)'(Y - Xb) \quad (3)$$

$$SST = Y'Y - \left(\frac{1}{N}\right)Y'JY \quad (4)$$

Where J is a nxn matrix.

Moreover, the variability of the Sum of Squares (SSR) varied, reflecting a p-1 degrees of freedom, suggesting that p was indicative of the number of predictor variables or parameters. Conversely, the Sum of Squares of Errors (SSE) was inversely related to the number of respondents, denoted by n, minus p. Lastly, the Sum of Squares of Totals (SST) adheres to the expected pattern, exhibiting a degree of freedom of n minus 1.

The subsequent equation represents the Analysis of Variance (ANOVA), alongside Mean Square Regression (MSR) and Mean Square Error (MSE). MSE, defined as a risk measure that encapsulates the expected value of a squared error loss or quadratic loss, is derived by averaging the squares of the deviations, which represent the error. This deviation is the difference between the estimated value obtained by the estimator and the value actually measured. Such disparities may arise from random error or a failure to incorporate relevant information that could lead to a more precise estimation. Table 2 delineates the Analysis of Variance (ANOVA). The equations are as follows:

$$MSR = \frac{SSR}{p-1} \quad (5)$$

$$MSE = \frac{SSE}{n-p} \quad (6)$$

Table 2 *Summary of ANOVA*

Source of Variation	SS	df	ms
Regression	$SSR = bX'Y - \left(\frac{1}{n}\right)Y'JY$	p-1	$MSR = \frac{SSR}{p-1}$
Error	$SSE = (Y - Xb)'(Y - Xb)$	n-p	$MSE = \frac{SSE}{n-p}$
Total	$SST = Y'Y - \left(\frac{1}{N}\right)Y'JY$	n-1	

3.3 Statistical Error Measurement

One technique for assessing the statical model's performance is rotation estimation. The analysis uses cross-validation. It is usually utilised in scenarios where the objective is anticipated, as well as for determining the approximate accuracy of a predictive model's performance in real-world scenarios. According to Kutner et al. (2005), the method that is frequently employed is mean square error (MSE).

Mean square error (MSE) is to measures the average of the squares of the errors and corresponding to the expected value of square error loss. $\hat{y}$ is the estimated value of a response variable in a linear regression model. The equation of MSE as in equation 7:

$$MSE = (y - \hat{y})^2 / N \quad (7)$$

4. Result and Discussion

4.1 Descriptive Analysis

Descriptive analysis serves as a method to evaluate and provide a succinct and fundamental summary of the study's samples and measurements. Consequently, the data of the study are examined by calculating the mean and central tendency to assess the properties of each item individually. Moreover, descriptive analysis through the Statistical Package for the Social Sciences (SPSS) is an outstanding approach to distinguish the distribution of means.

4.1.1 Descriptive Data for Ages

Table 3 presents a collection of descriptive statistical analysis for a cohort consisting of 569 patients diagnosed with breast cancer. The demographic spectrum of the sample spans from 25 to 86 years of age, with an average age of 45.1 years, accompanied by a standard deviation of 13.98 years. This suggests a broad distribution of ages among the patients, ranging from approximately 28 years, with the majority falling between the ages of 33 and 39. It is important to note that this dataset is representative of a specific group of breast cancer patients, and the conclusions drawn should not necessarily be extrapolated to the broader breast cancer patient population. Factors such as the age at which individuals are diagnosed or the specific subtype of breast cancer may influence the demographics of this particular group.

Table 3 *Mean of ages patients*

Total (N)	Mean	Standard Deviation
569	45.1353	13.97965

4.2 Assumptions of Regression Analysis

4.2.1 Normality

The Q-Q (Quantile-Quantile) graph reveals that the observed sizes of tumours do not conform to a normal distribution. The data exhibits a right-skewed distribution, suggesting an excess of tumours above the expected size. Conversely, it appears that tumours below the predicted growth size are less frequently observed. Moreover, the Q-Q graph suggests a positive correlation between the size of tumours and the predicted growth rate. This implies that tumours surpassing the predicted size tend to be detected more frequently compared to those below the expected size.

J. of Technology Management and Business Vol. 11 No. 2 (2024) p. 72-82

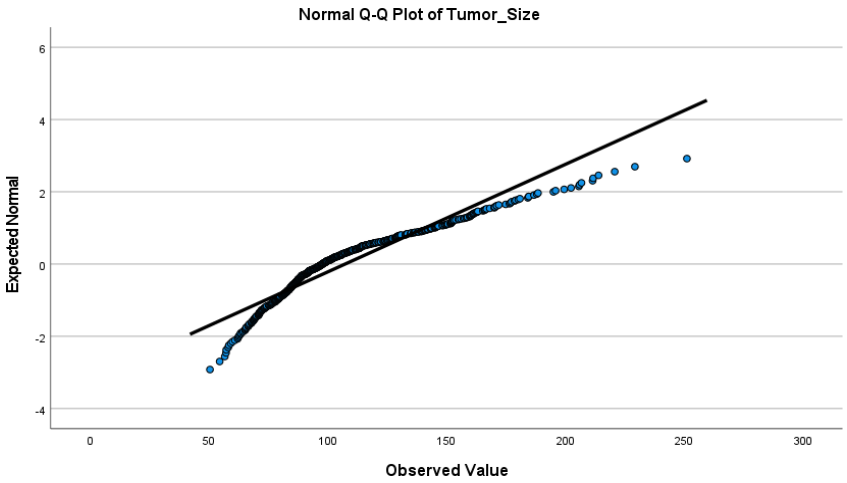

Fig. 3 *Q-Q plot of tumour size*

4.2.2 Variance of Residual

SPSS software was utilised to compute the residual variance. To determine the residual's condition variance, a scatter plot was employed. There appears to be no pattern to the dots, which suggests that the error terms have a zero mean. It may be inferred from this figure that the variance of the error terms is constant. This assumption is met. Table 4 displays the outcome of residual statistic.

Table 4 *Residual statistic*

Residuals Statistics					
	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	3.233	4.870	4.378	0.256	569
Residual	-1.533	1.585	0.000	0.391	569
Std. Predicted Value	-4.459	1.916	0.000	1.000	569
Std. Residual	-3.903	4.304	0.000	0.995	569

4.2.3 Multicollinearity Checking

Table 5 shows that all tolerance values are less than 0.99, all Variance Inflation Factor (VIF) values are less than 10, and the eigen values are decreasing from the highest to the lowest. The condition index has a value of 23.751, which is below 1000. This demonstrated that the multicollinearity condition was met and that the independent variables.

Table 5 *Coefficients of tolerance values and eigen values and explanation variances for data*

	Tolerance	VIF	Eigen value	Condition Index
(Constant)			3.974	1.000
Training and development	0.568	1.762	0.010	19.604
Performances appraisal	0.614	1.630	0.008	22.114
Selection and recruitment	0.634	1.577	0.007	23.751

4.3 Multiple Linear Regression

This research plans to utilize multiple linear regression analysis owing to the verification of three distinct conditions. This statistical technique aims to forecast the dependent variable's value in relation to various independent variables. According to the information presented in the table below, mass media exerts the most significant influence factor on the Radius of the breast cancer model, with a β value of 0.887, indicating a substantial impact on the radius.

This study successfully identifies significant relationships between the characteristics of tumours and their growth rate. It is observed that tumours with a larger radius demonstrate a strong positive correlation with an increased rate of development (beta = 0.887, $p < 0.001$), whereas tumours with a smoother texture exhibit a notable inverse relationship, suggesting a potential correlation with slower growth or diminished size (beta = -0.019, $p < 0.001$). The variable of compactness is shown to positively influence development (beta = 0.080, $p < 0.001$), suggesting that tumours with higher densities may exhibit accelerated growth. However, the precise extent of this relationship is not explicitly defined. The inference is drawn that enhanced cell-cell interactions and signalling mechanisms may play a role in explaining the observed patterns.

Table 6 *Multiple linear regression*

Model	β	Sigma Value
Age	0.06	0.147
Radius	0.887	< 0.001
Area	0.070	0.001
Smoothness	-0.19	< 0.001
Compactness	0.080	< 0.001
Concavity	0.009	0.240

All pertinent variables have been identified to affect the incidence of breast cancer. The model for predicting factors associated with breast cancer is presented below:

$\hat{Y} = 1.765 + 0.887$ Radius + 0.070 area – 0.19 smoothness + 0.080 compactness

4.3.1 Analysis of Variance (ANOVA)

The analysis also incorporates the use of ANOVA to ascertain the mean within a regression model and to verify the statistical significance of these findings. The mean error term has been determined to be 9.090. Furthermore, the F test statistic has achieved a P-value below 0.05, which affirms robust evidence in opposition to the null hypothesis.

Table 7 *Result of ANOVA*

Model	Sum of Squares	df	Mean Square	F-Value	P-Value
Regression	636237.505	6	106039.584	11664.981	0.000
Residual	5108.816	562	9.090		
Total	641346.321	568			

5. Conclusion

This study aimed to predict the high-risk factors of breast cancer in patients and to diminish future mortality rates. The analysis encompassed breast cancer patients from general hospital in Malaysia, and it was observed that a majority of these patients, between the ages of 33 and 39, exhibited colorectal cancer. Breast cancer is characterized by six determinants: age, radius, area, smoothness, compactness, and concavity. To explore the prediction of tumour size in breast cancer, employing multiple linear regression was necessary. The regression models identified significant factors such as age, radius, smoothness, and compactness. Research on breast cancer frequently applies a variety of statistical and computational methodologies aimed at predicting and comprehending the disease. Although machine learning (ML) and deep learning (DL) approaches are gaining prominence, the utilization of multiple linear regression continues to be advantageous in this field. This is attributed to its straightforwardness, interpretability, and particular benefits in specific situations.

The application of multiple linear regression in predicting high-risk factors of breast cancer while considering traits like age, radius, area, smoothness, compactness, and concavity offers valuable insights into the relationships

among these factors and tumour progression. The findings indicate a strong positive correlation between tumour size and radius, underscoring the significance of tumour dimensions in predicting growth rates. Moreover, a negative correlation with smoothness suggests a possible association with a slower growth rate or smaller tumour dimension. The significant correlation with compactness implies that tumours with a higher degree of compactness may exhibit accelerated growth. In conclusion, this methodology provides a comprehensive understanding of the multiple factors affecting tumour growth in breast cancer, thereby serving as an effective tool for predictive modelling and potentially aiding in clinical decision-making processes.

Acknowledgement

This research was supported by Universiti Tun Hussein Onn Malaysia (UTHM) through Multidisciplinary Research Grant (MDR) vot (Q701).

Conflict of Interest

The authors declare that there is no conflict of interest regarding the publication of the paper.

Author Contribution

The authors confirm contribution to the paper as follows: ***study conception and design:*** *Muhammad Ammar Shafi, Zulfana Lidinillah;* ***data collection:*** *Nor Faezan Abdul Rashid;* ***analysis and interpretation of results:*** *Muhammad Ammar Shafi, Zulfana Lidinillah, Shawal Iskandar, Mohd Saifullah Rusiman;* ***draft manuscript preparation:*** *Nur Ain Ebas, Mohd Saifullah Rusiman. All authors reviewed the results and approved the final version of the manuscript.*

References

Agresti, A. (1996). *An Introduction to Categorical Data Analysis Second Edition*.

Agresti, A. (1996). An introduction to categorical data analysis.

American Cancer Society. (2021). Breast Cancer Overview. Retrieved from https://www.cancer.org/cancer/breast-cancer/about.html

American Cancer Society. Cancer Facts & Figures 2014 Atlanta. (2014).

Basak, D., Pal, S., & Patranabis, D. C. (2007). Support Vector Regression. In *Neural Information Processing-Letters and Reviews* (Vol. 11, Issue 10).

Bednarek, A. K., Sahin, A., Brenner, A. J., Johnston, D. A., & Aldaz, C. M. (1997). Analysis of telomerase activity levels in breast cancer: positive detection at the in situ breast carcinoma stage. *Clinical Cancer Research : An Official Journal of the American Association for Cancer Research*, *3*(1), 11–16.

Bogdanova, N., Helbig, S., & Dörk, T. (2013). Hereditary breast cancer: ever more pieces to the polygenic puzzle. *Hereditary Cancer in Clinical Practice*, *11*(1), 12. https://doi.org/10.1186/1897-4287-11-12

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., & Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, *68*(6), 394-424. https://doi.org/10.3322/caac.21492

Bryman, A., & Bell, E. (2015). Business research methods (4th ed.). Oxford, UK: Oxford University Press.

Chen, M., Wu, X., Zhang, J., & Dong, E. (2021). Prediction of total hospital expenses of patients undergoing breast cancer surgery in Shanghai, China by comparing three models. BMC Health Services Research, 21(1), 1–9. https://doi.org/10.1186/s12913-021-07334-y

Draper, N. R., & Smith, H. (1998). Multiple regression applied to analysis of variance problems. Applied Regression Analysis, 473-504.

Hu, Z. (2020, August). Based on multiple linear regression analysis of ability of medical postgraduate students to express SPSS experiment results with three-line table. In *Journal of Physics: Conference Series* (Vol. 1592, No. 1, p. 012060). IOP Publishing. https://doi.org/10.1088/1742- 6596/1592/1/012060

Jacquillat, C., Weil, M., Baillet, F., Borel, C., Auclerc, G., De Maublanc, M. A., ... & Khayat, D. (1990). Results of neoadjuvant chemotherapy and radiation therapy in the breast-conserving treatment of 250 patients with all stages of infiltrative breast cancer. *Cancer*, *66*(1), 119-129.

Junaid, Muhammad. (2022). Application of Regression Analysis in Advance Research. 10.5281/zenodo.10722108.

Kutner H. Michael, Nachtsheim, Neter John and Li William. (2004).Applied Linear Statistical Models.Fifth Edition.

Lee, M. S., Ma'ruf, C. A. A., Izhar, D. P. N., Ishak, S. N., Jamaluddin, W. S. W., Ya'acob, S. N. M., & Kamaluddin, M. N. (2019). Awareness on breast cancer screening in Malaysia: a cross sectional study. *BioMedicine*, *9*(3).

Madala, S., Macdougall, K., Morvillo, G., Guarino, R., & Sokoloff, A. (2021). Guillain-Barré Syndrome as a Presenting Symptom in Breast Cancer: The Importance of Considering Paraneoplastic Neurologic Syndrome. Cureus, 13(9), 2–7. https://doi.org/10.7759/cureus.17932

Mariapun, S., Li, J., Yip, C. H., Miyake, T. M., Shi, J., Tamimi, R. M., . . . Teo, S. H. (2019). Pathology-guided system for personalized breast cancer prognosis. npj Breast Cancer, 5(1), 15. doi:10.1038/s41523-019-0103-1

Mayo Clinic. (2022). Breast Cancer Symptoms and Causes. https://www.mayoclinic.org/diseases-conditions/breast-cancer/symptoms-causes/syc- 20352470

National Cancer Institute. (2021). Breast cancer treatment (PDQ®)–patient version. Retrieved from https://www.cancer.gov/types/breast/patient/breast-treatment-pdq

National Cancer Registry, Ministry of Health Malaysia. (2010). Malaysian Cancer Statistics – Data and Figure Peninsular Malaysia.

Pencina, M. J., D'Agostino, R. B., & Vasan, R. S. (2008). Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond. Statistics in Medicine, 27(2), 157-172. https://doi.org/10.1002/sim.2929

Sarstedt, M., Mooi, E., Sarstedt, M., & Mooi, E. (2019). Regression analysis. *A concise guide to market research: The process, data, and methods using IBM SPSS Statistics*, 209-256.

Shafi, M. A., & Rusiman, M. S. (2015). The use of fuzzy linear regression models for tumor size in colorectal cancer in hospital of Malaysia. *Applied Mathematical Sciences*, *9*(56), 2749-2759.

Sykes, A. O. (1993). An introduction to regression analysis.

TechTarget. (2023, March 8). What is a botnet? SearchSecurity. Retrieved July 5, 2023, from https://searchsecurity.techtarget.com/definition/botnet

Wendy and Radzi. (2008). Editorial Cell Therapy Centre, Universiti Kebangsaan Malaysia Medical Centre.Med J Malaysia, (63), No 4.

World Health Organization. (2018). Latest global cancer data: Cancer burden rises to 18.1 million new cases and 9.6 million cancer deaths in 2018. *International agency for research on cancer. Geneva: World Health Organization*, 1-4.

World Health Organization. (2023). Publications of the World Health Organization are available on the WHO web site (www.who.int)

Xu, Y., Gong, M., Wang, Y., Yang, Y., Liu, S., & Zeng, Q. (2023). Global trends and forecasts of breast cancer incidence and deaths. *Scientific data*, *10*(1), 334.

Zheng, H. C., Zhou, J., Chen, Y. C., Yu, Y., Dai, W., Han, Y., ... & Jiang, S. F. (2023). The burden and trend of liver metastases in Shanghai, China: a population-based study. *European Journal of Cancer Prevention*, *32*(6), 517-524.