MSD

Medical Science and Discovery
ISSN: 2148-6832

Determination of risk factors using Nonlinear Principal Component Analysis in patients with breast tumour

Canan Demir[1]*

1 Van Yuzuncu Yil University, Vocational School of Health Services, Van, TR

*** Corresponding Author:** Canan Demir **E-mail:** canandemir@yyu.edu.tr

ABSTRACT

Objective: Breast cancer, which is the most common among women in the world and constitutes approximately 30% of all cancers, takes places near the top among the diseases that threaten women's health. The purpose of this study is to determine the risk factors in patients with breast tumours using nonlinear principal component analysis.

Materials and Methods: During the application process, a data set of 569 (357 benign, 212 malign) patients with breast tumours was used. To find independent features, the data set was reduced to two dimensions via nonlinear principal component analysis. The results were evaluated by comparing the success of the method with the ROC curve.

Results: The cut-off values for the radius, perimeter, area, smoothness and texture of the tumour were 14.19, 656.10, 0.09, 2.87 and 0.11, respectively. The sensitivity of the current values according to the results of ROC analysis was determined as 84% for radius, 80% for perimeter, 86% for the area and 94% for texture. It is seen that the method has an overall success of over 80% in detecting malignant tumours.

Conclusion: It is hoped that this method, which is used to reveal risk factors and identify distinctive features in breast tumours, will reduce medical costs and provide a second opinion to physicians. In terms of decision making, it is predicted that the method can recognize malignant tumours and reduce the need for unnecessary biopsy for benign tumours.

Keywords: Breast Tumours, Dimensional reduction, Nonlinear principal omponent analysis, Optimal Scalling

INTRODUCTION

Breast cancer occurs when cells in the breast divide and grow without reasonable control (1). Breast cancer mostly begins with the malfunction of the milk-producing ducts (invasive ductal carcinoma), and cancer cells can spread to lymph nodes and even to other parts of the body such as the lungs (2). Breast cancer, which ranks first among diseases that threaten women's health and constitutes approximately 23% of all cancers, is the most common type of cancer in women in the world (3, 4). The annual incidence of breast cancer is approximately 1.7 million cases in the world, with ~ 231,840 cases in the US and ~ 100,000 cases in Europe. Considering the risk of breast cancer, which has a substantial morbidity and mortality rate, especially in terms of women's health/life, and the early stage, effective treatment and good prognosis, the importance of implementing early diagnosis studies becomes clear (5, 6).

Nonlinear principal component analysis (NLPCA) is a descriptive dimension reduction method that provides numerical and visual results for data sets containing continuous, categorical or discrete variables with a linear or nonlinear relationship between them (7).

The aim of this study is to determine risk factors for patients with breast tumours using nonlinear principal component analysis.

Research Article

Received 12-07-2021

Accepted 23-07-2021

Available Online: 24-07-2021

Published 30-07-2021

MATERIAL and METHODS

In the study, from the free-access data site as application material (*http://mlr.cs.umass.edu/ml/machine-learningdatabase*. Access date: 04.05.2020) 11 variables data of the patient with breast tumour of 569 (357 benign, 212 malign) provided were used and the variables and their features are given in following.

The variables and their properties are as follow.

Radius: The radius of an individual nucleus is measured by averaging the length of the radial line segments defined by the centroid of the snake and the individual snake points.

Perimeter: The total distance between the snake points constitutes the nuclear perimeter.

Area: Nuclear area is measured simply by counting the number of pixels on the interior of the snake and adding one-half of the pixels in the perimeter.

Compactness: Perimeter and area are combined to give a measure of the compactness of the cell nuclei using the formula perimeter2/area.

Smoothness: The smoothness of a nuclear contour is quantified by measuring the difference between the length of a radial line and the mean length of the lines surrounding it.

Concavity: Chords are drawn between non-adjacent snake points, and how far the true boundary of the core extends within each chord is measured.

Concave Points: This feature is similar to Concavity but measures only the number, rather than the magnitude, of contour concavities.

Symmetry: The difference in length between lines perpendicular to the main axis is measured in both directions to the cell border.

Fractal Dimension: The perimeter of the nucleus is measured using increasingly larger 'rulers'. As the ruler size increases, with decreasing the precision of the measurement, the observed perimeter decreases. Plotting these to values on a log scale and measuring the downward slope gives (the negative of) an approximation to the fractal dimension.

Texture: The texture of the cell nucleus is measured by finding the variance of the gray scale intensities in the component pixels (8).

Methods

Breast tumour, which is common in many parts of the world, occurs in breast cells. Data of 569 patients with breast tumours, 212 malignant and 357 benign, were used for the study. This data was taken from the study conducted by Street et al. (8). The data set reached on 04.05.2020 was evaluated via NLPCA.

Nonlinear principal component analysis: The nonlinear principal component analysis aims to find x object scores and yj mean values in various ways under some limitations. Thus, the following function is minimized.

$$\sigma\,(X;Y) = n_w^{-1} \sum_j c^{-1}\, tr\left(\left(X - G_jY_j\right)' M_j\, W\left(X - G_jY_j\right)\right)$$

$$j = 1, \dots, m \quad (1)$$

Variance explanation rates for each dimension for multiple nominal variables;

$$VAF1_s = n_w^{-1} \sum_{j \epsilon J} v_j tr(Y_{js}'\, D_j\, Y_{js})$$

$$s = 1, \dots, p \quad (2)$$

For multiple non-nominal variables, it is calculated as follow;

$$VAF2_s = \sum_{j \notin J} v_j a_{js}^2$$

$$s = 1, \dots, p \quad (3)$$

Eigenvalues for each dimension are calculated by the following formula;

$$\sqrt{\lambda_s} = VAF1_s + VAF2_s$$

$$s = 1, \dots, p \quad (4)$$

And λ_s is the diagonal element of Λ. Total explained variance for multiple nominal and non-multiple nominal variables over the means of dimensions is calculated by the equation below;

$$tr\left(\sqrt{\Lambda}\right) = p^{-1} \sum_s VAF1_s + \sum_s VAF2_s$$

$$s = 1, \dots, p \quad (5)$$

This equation is known as total eigenvalues. Vector coordinates for NLPCA are calculated by the following equation;

$$VAF_{js} = v_j a_{js}^2 \quad s = 1, \dots, p \;\; \text{ve}$$

$$\text{j} \notin \text{J} \quad (6)$$

If the analysis is made for non-multiple variables, there are no missing observations or if it is determined passively, the correlation matrix is $q_j = G_j y_j \mathrm{R}$ and $R = n_w^{-1}\, Q'WQ$. The first p eigenvalue of R is equal to $\sqrt{\Lambda}$. If there are multiple nominal variables in the analysis, the p correlation matrices are calculated by equation 7;

$$R_s = n_w^{-1}\, Q_s' W Q_s$$

$$s = 1, \dots, p \quad (7)$$

It is calculated by equation 7. In Equation 7, q_{js} is calculated as $G_j y_j$ for multiple non-nominal variables and as

$$\frac{G_j Y_{js}}{\sqrt{Y'_{js}\ D_j Y_{js}}}$$

for multiple nominal variables (9).

The 1st eigenvalue of the R_s matrix is generally higher and is equal to $\sqrt{\lambda_s}$. Lower values of $\sqrt{\Lambda}$ usually belong to the 2nd and later eigenvalues of Rs. In calculating eigenvalues; if variable j is the complementary variable for the singular value decomposition of the R matrix, first the first column from the R matrix and j-th the row is removed, then R_{ij} is multiplied by $\sqrt{v_i v_j}$ (10).

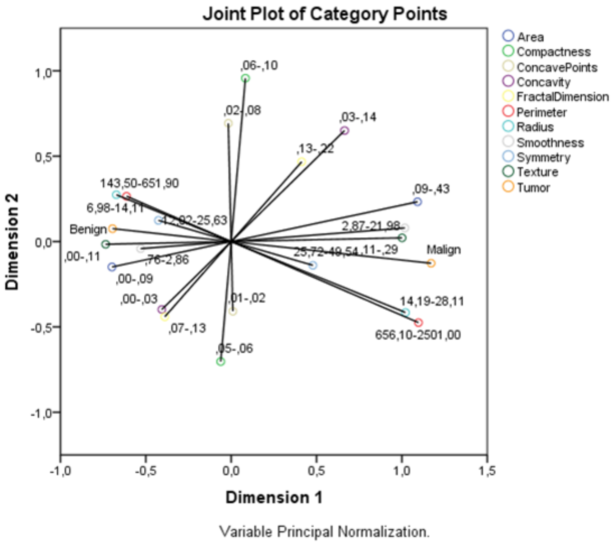
Receiver Operating Characteristics Curve Analysis: One of the common methods used to distinguish patients and healthy individuals by finding a cut-off point to determine the performance of continuous variables as a diagnostic test is the ROC (Receiver Operating Characteristics) curve. The ROC curve is the curve obtained by taking the measured values of the continuous variable (respectively) as the cut-off point, plotting the Sensitivity values on the Y-axis, and 1-Specificity values on the X-axis. The total area remaining under the curve is "1". If the area under the curve is 0.50 then the feature has no discriminating power, "1" indicates that it is 100%. The ROC curve summarizes the accuracy of the test with a single numerical value.

RESULTS

As a result of the applied principal components analysis, the results of the first two main components are given in **Table 1**. As seen in **Table 1**; 44.095% of the total variance was explained with the first main component, 15.686% with the second main component. Thus, eleven original variables have been reduced to two (basic) components that explain 59,781% of the total variance. When the principal component loads that express the correlations between original variables and principal components are examined; a high correlation between radius, perimeter, area, smoothness, texture, and tumour variables and the first major component; A moderate correlation was found with concave, symmetry, and fractal dimension. The contribution of compactness and concave point variables to the first fundamental component is almost negligible. Compactness is the variable that contributes the most to the second principal component. When the correlations between the variables analyzed with NLPCA are examined, there is a high correlation between the tumour and the variables of radius, perimeter, area, smoothness and texture; There is a moderate correlation between the tumour and the variables of concave, symmetry, and fractal size. No relationship was found between tumour and the variables of compactness and concave points. Parallel to the increase in the contribution of the categories to the dimensions and the increasing power of separation, the coefficient values of the dimensions also increase. In other words, moving away from the origin of the values of any category in the dimensions indicates that the effect of the said category in determining the size is higher. Accordingly, the "malign" category of the tumour variable with a value of 1.171 in the first dimension, the "0.06-0.10" category of the compactness variable with a value of 0.957 in the second dimension received the highest positive value, while in the first dimension with -0.735 the texture variable "0.00-0.11" category, “0.05-0.06” category of compactness variable has the highest negative value with -0.703 value.

In Figure 1, it is seen that the "malignant category" of tumour variable is highly associated with the "14.19 - 28.11" category of radius variable, "656.10-2501.00" category of perimeter variable, "0.09-0.43" category of area variable, "2.87-21.98" category of smoothness variable, and "0.11-0.29” category of texture variable, and it is moderately positively associated with the “0.03-0.14” category of concavity variable, “25.72-49.54” category the symmetry variable, “0.13-0.22” category of fractal dimension variable. Similarly, it was determined that the benign category of the tumour variable and the “6,98-14,11” category of the Radius variable, the “143.50-651.90” category of the perimeter variable, the “0.00-0.09” category of the area variable, the “0.76-2.86” category of the smoothness variable, and the “0.00-0.11” category of the texture variable were positively correlated.

Figure 1. Joint Plot of Category Points



It can be said that the categories that are close to the origin have low effects and they have no relation with other categories. In this context, it was observed that the effect of the concave points variable is very low and not related to other variables (**Figure 1**).

According to the ROC analysis results in the study; the area under the curve was found as 0.938 ± 0.010 for radius, perimeter and area, 0.876 ± 0.015 for smoothness, and 0.967 ± 0.007 for texture. Cut-off values for radius, perimeter, area, smoothness and texture are respectively seen as; 14.1950 (Sensitivity 84.4%, Specificity 87.1%), 656.25 (Sensitivity 80.7%, Specificity 91%), 0.0905 (Sensitivity 86.3%, Specificity 89.4%), 2.8780 (Sensitivity 71.2%, Specificity 88.5%), 0.1102 (Sensitivity 94.3%, Specificity 85.2%) (Table 2).

Table 1. Analysis results for the first two principal components

	Total (Vector Coordinates)		
	Dimension		
	1	2	Total
Tumour	0,814	0,010	0,824
Radius	0,687	0,114	0,801
Perimeter	0,672	0,126	0,798
Area	0,762	0,035	0,797
Compactness	0,005	0,673	0,678
Smoothness	0,539	0,003	0,542
Concavity	0,270	0,259	0,529
ConcavePoints	0,000	0,283	0,283
Symmetry	0,204	0,017	0,221
FractalDimension	0,161	0,205	0,366
Texture	0,736	0,000	0,736
Active Total	4,850	1,725	6,576
% of Variance	44,095	15,686	59,781

Table 2. ROC analysis summary

	Group	Cut-Off Value	Area Under The Curve	St. Error	Sensitivity	Specificity	*P*
Radius	Malign-Benign	14.1950	0.938	0,010	0.844	0.871	0,001
Perimeter	Malign-Benign	656.25	0.938	0,010	0.807	0.910	0,001
Area	Malign-Benign	0.0905	0.938	0,010	0.863	0.894	0,001
Smoothness	Malign-Benign	2.8780	0.876	0.015	0.712	0.885	0,001
Texture	Malign-Benign	0.1102	0.967	0.007	0.943	0.852	0,001

DISCUSSION

Breast cancer is a very common cancer; It has the second-highest incidence rate worldwide among all types of cancer and is ranked as the fifth leading cause of cancer-related death (11). Various risk factors have been identified for breast cancer. Sun et al. listed these risk factors as an agent, family history, reproductive factors, estrogen, and lifestyle, respectively. In their studies where they emphasized the importance of preventing breast cancer, they stated that current prevention methods such as screening, chemoprevention and biological prevention are more accurate and more effective than previous methods (12). Stating that the risk of breast cancer is lower in breastfeeding women than other women, Turkoz et al. pointed out that obesity and overweight may also be considered as risk factors at later ages (13). Early diagnosis is the cornerstone of preventing mortality in breast cancer (12). Therefore, the importance of imaging for the detection and diagnosis of breast cancer cannot be denied. The chance of treating cancer depends primarily on early diagnosis, and treatment choice depends on the level of malignancy. For this reason, it is very important to detect cancer, to separate cancerous from benign and healthy ones, and determine the level of malignancy. The geometric organization of cells in tissue can affect proliferation, propagation, branching, stem cell properties and cancer cell survival and invasion (14). Traditionally, pathologists use histopathological images of biopsy samples taken from patients, examine them under a microscope, and make decisions based on personal experience. However, these decisions are subjective and often lead to variability (15).

Grove et al. in their study, in which they stated that tumour shape and intratumoral density variation reflect tumour biology and may affect patient survival, they show that quantitative imaging biomarkers can be used as an additional diagnostic tool in the treatment of lung adenocarcinomas (16).

In a study, it was stated that the variation in the size and shape of the tumour can be used as an indicator of the presence of cancer (17). Using the Bayesian network inference approach, Hussain et al. found significant associations between morphological features extracted from prostate cancer images (18).

Computer-aided diagnostic systems based on tissue and morphological analysis have proven to be extremely sensitive in evaluating breast tumours. Zhou et al. were able to distinguish benign breast tumours with high accuracy and short training time in their study (19).

In a standard principal component analysis, the aim is to find fewer new variables consisting of combinations of these variables that can explain the total variance of the original variables as much as possible (20). The size reduction feature of principal components analysis has been used in this article. The risk factors affecting the malignancy of the tumour were determined by reducing the data set into two dimensions. Accordingly, the radius, perimeter, area, softness and tissue of the tumour were found to be significantly effective on malignancy. In many studies, it has been determined that the radius, perimeter and area of the tumour are effective. However, in this study, it has been shown that the above-mentioned characteristics with high effects on the malignancy of the tumour increase the malignancy after which values.

It can be said that the tumour can now be a malignant tumour when the radius of the tumour is 14.19, the perimeter is 656.10, the area is 0.09, the smoothness is 2.87 and the texture is above 0.11. The sensitivity of the current values according to the results of ROC analysis was determined as 84% for radius, 80% for perimeter, 86% for the area and 94% for texture. It is seen that the method has an overall success of over 80% in detecting malignant tumours.

In particular, radius, perimeter, area and texture variables can be shown as important risk factors, but it can be said that the variables of compactness, concavity, concave points, symmetry and fractal dimension have a low effect on malignancy signs.

CONCLUSIONS

The results show that the use of morphological features is effective and safe. Determining the morphological features and risk factors in breast tumours can be seen as an advantage. In terms of decision making, it is predicted that the method can recognize malignant tumours and reduce the need for unnecessary biopsy for benign tumours. It has been suggested that breast tumours start early in life and are shaped by the number of cells at risk, the integrity of these cells and the environment they are exposed to (**21**). Therefore, it is hoped that this method, which is used to reveal risk factors and distinguish features in breast tumours, will reduce medical costs and provide a second opinion to physicians.

Acknowledgments: None

Author Contributions: CD: Data collection, Formal analysis, Methodology, Project administration, Statistical Analyses, **CD:** Article writing and revisions

Financial & competing interest's disclosure: The authors have no relevant affiliations or financial involvement with any organisation or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript. This includes employment, consultancies, honoraria, stock ownership or options, expert testimony, grants or patents received or pending, or royalties.

Conflict of interest: The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article. This research did not receive and specific grant from funding agencies in the public, commercial or not-for-profit sectors.

REFERENCES

1. Mambou SJ, Maresova P, Krejcar O, Selamat A, Kuca K. Breast cancer detection using infrared thermal imaging and a deep learning model. Sensors-Basel 2018;18(9):2799.
2. Gardezi SJS, Elazab A, Lei B, Wang T. Breast cancer detection and diagnosis using mammographic data: systematic review. J Med Internet Res 2019;21(7):e14464.
3. Usmani A, Lateef M. Evaluation of C-reactive protein in breast cancer by enzyme linked immunoassay technique. J Pak Med Assoc 2021;71:424-8.
4. Irfan R, Memon H, Umrani IN, Soomro H. Breast cancer awareness among pharmacy and physiotherapy students of medical university Nawabshah. J Pak Med Assoc 2021;71:297-301.
5. Coughlin SS. Epidemiology of Breast Cancer in Women. Adv Exp Med Biol 1152, Springer Nature Switzerland AG; 2019, pp 9-29.
6. Carioli G, Malvezzi M, Rodriguez T, Bertuccio P, Negri E, Vecchia CL. Trends and predictions to 2020 in breast cancer mortality in Europe. The Breast 2017;36:89-95.
7. Demir C. Dimensionality Reduction Technique: Principal Component Analysis. Current Researches and New Trends IVPE; 2020, pp 12-17.
8. Street WN, Wolberg WH, Mangasarian OL. Nuclear feature extraction for breast tumor diagnosis. International Symposium Electronic Imaging: Science and Technology; 1993 Feb 1-4; San Jose, CA, USA. vol. 1905, pp. 861-870.
9. Demir C, Keskin S. Artificial neural network approach for nonlinear principal components analysis. Int J Curr Res 2021;13(1)15987-92.
10. Ferrari PA, Barbiero A. Nonlinear Principal Component Analysis. Modern Analysis of Customer Surveys, eds R. S. Kenett and S. Salini (Chichester: John Wiley and Sons, Ltd.) 2011, pp 333–56.
11. Ahn HR, Kang SY, Youn HJ, Jung SH. Hyperglycemia during Adjuvant Chemotherapy as a Prognostic Factor in Breast Cancer Patients without Diabetes. J Breast Cancer 2020;23(4):398-409.
12. Sun YS, Zhao Z, Yang ZN, Xu F, Lu HJ, Zhu ZY, et al. Risk factors and preventions of breast cancer. Int J Biol Sci 2017;13(11):1387-97.
13. Turkoz FP, Solak M, Petekkaya I, Keskin O, Kertmen N, Sarici F, et al. Association between common risk factors and molecular subtypes in breast cancer patients. The Breast 2013;22(3):344-50.
14. Lee J, Abdeen AA, Wycislo KL, Fan TM, Kilian KA. Interfacial geometry dictates cancer cell tumorigenicity. Nat mater 2016;15(8):856-62.
15. Kumar R, Srivastava R, Srivastava S. Detection and Classification of Cancer from Microsc opic Biopsy Images Using Clinically Significant and Biologically Interpretable Features. J Med Eng 2015; 1-15.
16. Grove O, Berglund AE, Schabath MB, Aerts HJ, Dekker A, Wang H, et al. Quantitative computed tomographic descriptors associate tumor shape complexity and intratumor heterogeneity with prognosis in lung adenocarcinoma. PloS one. 2015;10(3):e0118261.
17. Teramoto A, Tsukamoto T, Kiriyama Y, Fujita H. Automated Classification of Lung Cancer Types from Cytological Images Using Deep Convolutional Neural Networks. Biomed Res Int 2017; 1-6.
18. Hussain L, Ali A, Rathore S, Saeed S, Idris A, Usman MU, et al. Applying bayesian network approach to determine the association between morphological features extracted from prostate cancer images. Ieee Access 2018;7:1586-601.
19. Zhou S, Shi J, Zhu J, Cai Y, Wang R. Shearlet-based texture feature extraction for classification of breast tumor in ultrasound image. Biomed Signal Process Control 2013; 8(6): 688-96.
20. Kumar BLNP, Prabukumar M. Hyperspectral image classification using fuzzy-embedded hyperbolic sigmoid nonlinear principal component and weighted least squares approach. J. Appl. Remote Sens 2020; 14(2) 024501.
21. Barnard ME, Boeke CE, Tamimi RM. Established breast cancer risk factors and risk of intrinsic tumor subtypes. Biochimica et Biophysica Acta (BBA)-Reviews on Cancer. 2015;1856(1):73-85.