Kernel PCA based on Hotelling Multivariate Control Chart for Monitoring Breast Cancer Diagnostic

Aurell Faza Ashilla, Awang Putra Sembada R, I Melda Puspita Loka, Sukma Adi Perdana, and Muhammad Ahsan

Abstract: **Breast cancer is one of the diseases that is a scourge for women. The diagnosis of cancer is divided into benign breast cancer and malignant breast cancer. Accuracy in determining the diagnosis of breast cancer can help in patient treatment. To monitor breast cancer diagnoses, this article proposed a Kernel PCA based Hotelling's T^2 Multivariate control chart. All features that are believed to affect cancer diagnosis are reduced with the Kernel PCA to overcome multicollinearity. The Kernel PCA based Hotelling's T^2 The multivariate chart employs the bootstrap approach, a nonparametric resampling method to estimate the control limit. A study was conducted to compare the performance of the control charts with logistic regression to see the superiority of the control chart in diagnosing the type of breast cancer. The accuracy of Kernel PCA based Hotelling's T^2 The multivariate graph is 89.63%. The logistic regression performance is better at classifying breast cancer diagnoses compared to Hotelling's T^2 since it has a bigger accuracy. These results make sense because the function of logistic regression is to classify. Whereas in Hotelling's T^2, we use the concept of in-control and out of control. However, to predict the diagnosis of breast cancer, the performance of Hotelling's T^2 with an accuracy value close to 90%, can be said to be good.**

***Index Terms*—breast cancer, control chart, hotelling's T^2, kernel PCA**

I. Introduction

ONE of the diseases that is a scourge for women is breast cancer. Based on the data from the American Cancer Society in 2017, more than 252,710 women with breast cancer have been diagnosed and approximately 16% of them are losing their lives because of this disease. Like other cancers, breast cancer can also divide into benign breast cancer and malignant breast cancer. Accuracy in determining benign or malignant breast cancer can help with treatment for breast cancer patients. Advances in medical technology and information disclosure have made many medical datasets available. One of them is about breast cancer from the UCI machine learning repository and collected from the University of California, Irvine. These data can be equipment to develop new methods for determining benign or malignant breast cancer.

PCA (Principal Component Analysis) is a multivariate technique that can be used to reduce variables. The PCA method is very useful for data with many variables and has a correlation between its variables. The purpose of PCA is to reduce the variables to fewer variables without losing the information contained in the original data. Hence, PCA is appropriate to be implemented for breast cancer data from the UCI machine learning repository. It is because breast cancer data from the UCI machine learning repository have ten features, such as radius (mean distances from center to points on the perimeter, texture (standard deviation of grayscale values), perimeter, area, smoothness (local variation in radius lengths), compactness (perimeter 2 / area – 1.0), concavity (severity of concave portions of the contour), concave points (number of concave portions of the contour), symmetry, and fractal dimension ("coastline approximation" – 1).

The control chart is a technique that uses graphics to monitor the quality of a manufacturing process. The main purpose of the control chart is to find assignable causes of process variation. Assignable causes are caused by factors that are not part of the process. When assignable causes are detected, then the process is out of control. A control chart consists of three horizontal lines such as the lower control limit (LCL), the centerline, and the upper control limit (UCL). A process is considered in control when the data point falls between UCL and LCL. A data point that falls outside the control area is indicated as an out-control signal. In general, control charts are effective tools to eliminate process variability and estimate process parameters [1]. One of the much-used control charts is Hotelling's T^2. Hotelling's T^2 assumes the data follow a multivariate normal distribution [2]. If the data do not follow a multivariate normal distribution, then the control limit of Hotelling's T^2 can be calculated using the bootstrap approach.

There is much research regarding detecting or diagnosing phenomena using control charts. Das and Sugal conducted research on the identification of hot and cold spots in the

Manuscript received November 24, 2023; revised April 18, 2024. This work was supported by Institut Teknologi Sepuluh Nopember under the project scheme of the Publication Writing and IPR Incentive Program (PPHKI) 2024.

A. F. Ashilla is magister student in Department of Statistics, Institut Teknologi Sepuluh Nopember, Surabaya, 60111 Indonesia (e-mail: aurellashilla@gmail.com).

A. P. Sembada is magister student in Department of Statistics, Institut Teknologi Sepuluh Nopember, Surabaya, 60111 Indonesia (e-mail: awangputrasembadar1@gmail.com).

I M. P. Loka is a magister student in Department of Statistics, Institut Teknologi Sepuluh Nopember, Surabaya, 60111 Indonesia (e-mail: imeldapuspitaloka@gmail.com).

S. A. Perdana is doctoral student in Department of Statistics, Institut Teknologi Sepuluh Nopember, Surabaya, 60111 Indonesia (e-mail: sukma_adi@stainkepri.ac.id).

M. Ahsan is Assistant Profesor in Department of Statistics, Institut Teknologi Sepuluh Nopember, Surabaya, 60111 Indonesia (e-mail: muh.ahsan@its.ac.id).

genome of Mycobacterium tuberculosis using Shewhart control charts [3]. Ahsan et al. created an intrusion detection system using the bootstrap resampling approach of Hotelling's T^2 control chart based on successive difference covariance matrix [4]. Rogerson and Yamada [5] proposed a multivariate cumulative sum approach to detect changes in spatial patterns and applied it to county-level breast cancer data in the Northeastern United States. The results of the comparison suggested that the multivariate chart performed well.

Some control charts have the assumption that they rely on normal distribution to establish control limits. The bootstrap method is a nonparametric technique that can be used to set control limits without considering the parametric distribution assumptions of the observed data [6]–[8]. Taking into account the effectiveness of the control chart to detect and diagnose, we propose the use of the PCA control chart [9] to diagnose the type of breast cancer. Furthermore, the performance of the PCA control chart was compared with logistic regression to see the superiority of the control chart in diagnosing the type of breast cancer.

II. MATERIAL AND METHOD

A. Principal Component Analysis

Principal Component Analysis (PCA) is a method to significantly reduce variables. PCA can reduce the dimensions of the observational data but does not lose significant information in the data. The problem with using PCA is to find the eigenvalues and eigenvectors [7].

PCA is a basis transformation to diagonalize an estimate of the covariance matrix of the data x_k, $k=1,...,\ell$, $x_k \in \square^N$, $\sum_{k=1}^{\ell} x_k = 0$ defined as

$$C = \frac{1}{\ell}\sum_{j=1}^{\ell} x_j x_j^T \tag{1}$$

The principal component is referred to as the new coordinates on the basis of the eigenvectors, which are orthogonal projections to the eigenvectors.

B. Kernel PCA

Assume for now that the data is mapped to a centralized feature space $\Phi(x_1),..,\Phi(x_\ell)$, i.e., $\sum_{k=1}^{\ell}\Phi(x_k)=0$. To perform PCA for the covariance matrix

$$\overline{C} = \frac{1}{\ell}\sum_{j=1}^{\ell}\Phi\left(x_j\right)\Phi\left(x_j\right)^T \tag{2}$$

Find the eigenvalues $\lambda \geq 0$ and eigenvectors $V \in F \setminus \{0\}$ that satisfy $\lambda V = \overline{C}V$. Substituting, note that all solutions V lies in the range $\Phi(x_1),..,\Phi(x_\ell)$. This implies that an equivalent system can be considered.

$$\lambda\left(\Phi\left(x_k\right)\cdot V\right) = \left(\Phi\left(X_K\right)\cdot \overline{C}V\right) \text{ for all } k=1,...,\ell \tag{3}$$

And there exist coefficients $\alpha_1,...,\alpha_\ell$ such that

$$V = \sum_{i=1}^{\ell}\alpha_i\Phi\left(x_i\right) \tag{4}$$

by substitution and defined the matrix with an $\ell x \ell$ matrix *K* by

$$K_{ij} := \left(\Phi\left(x_i\right)\Phi\left(x_j\right)\right) \tag{5}$$

obtained

$$\ell\lambda\alpha = K\alpha \tag{6}$$

for non-zero eigenvalues.

The normalized solution α^k belongs to the non-zero eigenvalues by requiring the corresponding vector F to be normalized, i.e., $\left(V^k\cdot V^k\right)=1$ obtained.

$$1 = \sum_{i,j=1}^{\ell}\alpha_i^k\alpha_j^k\left(\Phi\left(x_i\right)\cdot\Phi\left(x_j\right)\right) = \left(\alpha^k\cdot K\alpha^k\right) = \lambda_k\left(\alpha^k\cdot\alpha^k\right) \tag{7}$$

For principal component extraction, the projection of the test point image $\Phi(x)$ to the eigenvectors V^k in *F* is calculated according to

$$\left(V^k\cdot\Phi(x)\right) = \sum_{i=1}^{\ell}\alpha_i^k\left(\Phi\left(x_i\right)\cdot\Phi(x)\right) \tag{8}$$

In the $\Phi\left(x_i\right)$ explicit form they are only needed in the dot product. Therefore, it is possible to use a kernel function to calculate the product of this point without actually running the map Φ [10] for multiple kernel choices $k(x,y)$, it can be shown by the method of functional analysis that there is a map Φ into some point product space *F* (possibly infinite dimensions) so calculate *k* the product of the point in *F*. Kernels that have been used successfully in support vector machines [11]–[13] include the polynomial kernel.

$$k(x,y) = (x\cdot y)^d \tag{9}$$

Radial basis functions $k(x,y) = \exp\left(-\|x-y\|^2/\left(2\sigma^2\right)\right)$, and sigmoid kernel $k(x,y) = \tanh\left(\kappa(x,y)+\Theta\right)$. It can be shown that polynomial kernels of degree d correspond to a map Φ into a feature space which is spanned by all products of d entries of an input pattern, e.g., for the case of $N=2, d=2$ [6].

$$\left(x\cdot y\right)^2 = \left(x_1^2, x_1x_2, x_2x_1, x_2^2\right)\left(y_1^2, y_1y_2, y_2y_1, y_2^2\right)^T \tag{10}$$

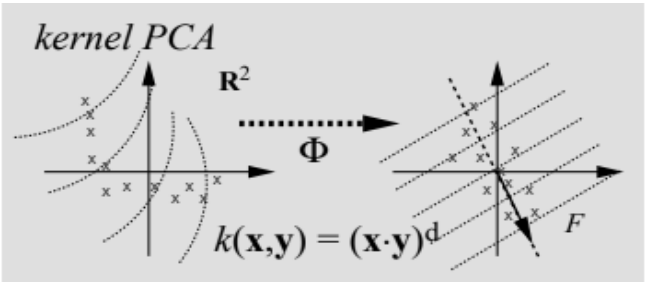

Fig. 1. Basic idea of Kernel PCA: by using a non-linear kernel function

Figure 1 suggests that we can analyze the data using all possible combinations of pixel values (d-th order products). This approach, when combined with a kernel function, captures more complex relationships between the data points compared to standard PCA for all occurrences of $\left(\Phi(x)\Phi(y)\right)$.

C. The Hotelling's T^2 control chart

The most common multivariate quality control methods are based on Hotelling's T^2 statistics. This T^2 statistic is

equivalent to the square of the Mahalanobis distance, and the calculation is based on the classical sample mean vector and the classical sample variance-covariance matrix. In Phase I, historical data sets from observations are analyzed to determine whether a process is in control and to estimate process parameters under control, control limits, and to identify and eliminate multivariate outliers. In phase II, estimates and control limits are used to examine data obtained during the industrial process to detect deviations from the parameter estimates.

The Hotelling's T^2 control chart is a widely used statistical tool and is considered the most common multivariate control chart to monitor multivariate variability in industrial processes. let $\mathbf{X}=\left(X_{ij1},,,X_{ij2},,,...,,,X_{ijp}\right)^t$ to show a vector *p* x 1 that represents the characteristic *p* of the *j*-th observations in the *i*-th subgroup, *i* = 1, 2,. ., *m* and *j* = 1, 2,. . , *n*, where *n* is the size of the subgroup and m is the number of subgroups. assume that $X_{ij}{}'s$ is independent and identically distributed and follows a multivariate normal distribution with the mean vector $\boldsymbol{\mu}$ and the variance-covariance matrix $\boldsymbol{\Sigma}$ when the process is in control. If the value of the process parameter is unknown, data from the initial m subgroups are collected when the process is believed to be in control. Then, the unbiased estimates of the mean vector $\boldsymbol{\mu}$ and the variance-covariance matrix $\boldsymbol{\Sigma}$, respectively, are given by the following.

$$\bar{\bar{x}}=\frac{1}{m}\sum_{i=1}^{m}\bar{X}_i \text{ and } \bar{S}=\frac{1}{m}\sum_{j=1}^{m}S_i \tag{11}$$

where $\bar{\mathbf{X}}_i$ denotes the mean vector for the *i*-th subgroup, and $\mathbf{S}_i$ denotes the unbiased estimate of the variance-covariance matrix for the *i*-th subgroup. That is,

$$\bar{X}_i=\frac{1}{n}\sum_{j=1}^{n}X_{ij} \text{ and } S_i=\frac{1}{n-1}\sum_{j=1}^{n}\left(X_{ij}-\bar{X}_i\right)\left(X_{ij}-\bar{X}_i\right)^t \tag{12}$$

The Hotelling's T^2 control chart is constructed using these estimated parameters [14]. As mentioned above, the control chart is first used to retrospectively test whether the process was in control when initial subgroups of m were drawn (Phase I). After the initial control has been established, the control chart can be used to monitor the process online, that is, the values of subgroup averages are plotted one at a time on the chart as each new subgroup is obtained (Phase II). In this paper, we consider phase I. The statistic plotted on the Hotelling's T^2 control chart for each initial subgroup is calculated as follows.

$$T_i^2=n\left(\bar{x}_i-\bar{\bar{x}}\right)^t\bar{S}^{-1}\left(\bar{x}_i-\bar{\bar{x}}\right),\ i=1,2,3...,m \tag{13}$$

The UCL of this control chart is given as follows.

$$UCL_{T^2}=\frac{p\left(m-1\right)\left(n-1\right)}{\left(mn-m-p+1\right)}F_{p,mn-m-p+1} \tag{14}$$

where $F_{v1,v2,\alpha}$ is the $\left(1-\alpha\right)^{th}$ percentile point of the *F* distribution with v_1 and v_2 degrees of freedom, and α is the desired false alarm probability for each subgroup. The lower control limit (LCL) is usually set to zero.

D. Logistic regression

Assume that there is one independent variable, *X,* and one dependent variable, *Y,* that have two categories. Let $\left(x\right)=P\left(Y=1\middle|X=x\right)=1-P\left(Y=0\middle|X=x\right)$. The logistic regression model can be written as follows.

$$\pi\left(x\right)=\frac{\exp\left(\beta_0+\beta_1X_1\right)}{1+\exp\left(\beta_0+\beta_1X_1\right)} \tag{15}$$

The extended model applies to multiple binary logistic regression.

$$\pi\left(x_i\right)=\frac{\exp\left(\beta_0+\beta_1X_1+\beta_2X_2+...+\beta_kX_k\right)}{1+\exp\left(\beta_0+\beta_1X_1+\beta_2X_2+...+\beta_kX_k\right)} \tag{16}$$

The $\beta_0,\beta_1,...,\beta_k$ are the parameters for the model. The estimation for the parameters is determined by the maximum likelihood estimation. The first step to estimate the parameter is to define the likelihood function. Assume there are $Y_1,Y_2,...,Y_N$ binomial random variables. As the observations are assumed to be independent, the likelihood function for these binomial random variables can be seen in the following formula.

$$L\left(\beta\right)=\prod_{i=1}^{N}\frac{n_i!}{y_i!\left(n_i-y_i\right)!}\pi\left(x_i\right)^{y_i}\left(1-\pi\left(x_i\right)\right)^{n_i-y_i} \tag{17}$$

$$L\left(\beta\right)=\prod_{i=1}^{N}\frac{n_i!}{y_i!\left(n_i-y_i\right)!}\pi\left(x_i\right)^{y_i}\left(1-\pi\left(x_i\right)\right)^{n_i-y_i} \tag{18}$$

Taking the natural logarithm of $L\left(\beta\right)$,

$$Ln\left(L\left(\beta\right)\right)=\sum_{i=1}^{N}y_iLn\left[\pi\left(x_i\right)\right]-\left(1-y_i\right)\ln\left[1-\pi\left(x_i\right)\right] \tag{19}$$

Differentiating $L\left(\beta\right)$ with respect to $\beta_0,\beta_1,...,\beta_k$. Set the result of this differentiation equal to zero. This result is not a closed form formula, so we require iterative methods to get the estimated coefficients, for example, the Iterative Weighted Least Squares method.

E. Breast Cancer

Breast cancer is one of the most common cancers in women worldwide, accounting for about 570,000 deaths in 2015. More than 1.5 million women (25% of all women with cancer) are diagnosed with breast cancer every year worldwide. In America, an estimated 30% of all new cancer cases (252,710) among women were breast cancer in 2017 [15]. Early diagnosis of the disease can lead to a good prognosis and a high survival rate. Mammography is a screening approach that is widely used in the detection of breast cancer and has been shown to help reduce mortality effectively. Other screening methods, such as magnetic resonance imaging (MRI), which are more sensitive than mammography, have also been applied and studied over the last decade. Although the incidence rate of breast cancer in America is increasing from year to year, the mortality rate is decreasing due to widespread early detection and continued medical therapy. There are many risk factors such as gender, aging, estrogen, family history, gene mutations, and an unhealthy lifestyle, which can increase the chances of getting breast cancer. Biological therapies have been developed in recent years and have been shown to be beneficial for breast cancer.

The mortality rate from breast cancer is higher in developing countries than in developed countries. The main cause of the increase in cancer deaths in developing countries is the lack of effective screening programs that can detect conditions before cancer, as well as detect cancer at an early stage so that treatment is carried out before cancer is at an advanced stage. In addition to the lack of a screening program, there is also a lack of knowledge, ability, and access for treatment. Therefore, early breast self-examination and education on appropriate treatment for the community are very necessary. Appropriate public knowledge about breast cancer and early detection efforts is still lacking. There is a need to increase understanding of breast cancer, since early diagnosis and surgery will increase the chances of cure and increase life expectancy. Thus, in the end, it can reduce morbidity and mortality and improve the quality of life of patients with breast cancer.

III. DATASET

The data set used is a Wisconsin Breast Cancer Data Set from the University of Wisconsin Hospital to check the accuracy of the performance and evaluate the proposed method. The target variable is diagnostic, and the rest are predictor/feature variables. The features are computed from a digital image of a fine needle aspirate (FNA) of a breast mass. Information about the data set is shown in Table 1.

TABLE I
DATASET INFORMATION

Variable	Detail
Diagnosis (Y)	The diagnosis of breast tissues (1 = malignant, 0 = benign)
radius_mean (X1)	mean of distances from center to points on the perimeter
texture_mean (X2)	standard deviation of gray-scale values
perimeter_mean (X3)	mean size of the core tumor
area_mean (X4)	mean of local variation in radius lengths
smoothness_mean (X5)	
compactness_mean (X6)	mean of perimeter^2 / area - 1.0
concavity_mean (X7)	mean of severity of concave portions of the contour
concave points_mean (X8)	mean for number of concave portions of the contour
symmetry_mean (X9)	
fractal_dimension_mean (X10)	

IV. RESULT AND DISCUSSION

A. Principal Component Analysis

The data set used is a data set of breast cancer patients with malignant and benign tumor types, which are tumor types, from the Wisconsin Diagnostic Breast Cancer data set. Hotelling's T^2 Control Chart and Logistic Regression is used to predict whether the patient is a malignant or benign tumor.

The following is the correlation matrix of the predictor variables as follows.

	X1	X2	X3	X4	X5	X6	X7	X8	X9
X2	0.32								
X3	0.99	0.33							
X4	0.98	0.32	0.98						
X5	0.17	-0.02	0.20	0.17					
X6	0.51	0.23	0.55	0.49	0.65				
X7	0.67	0.30	0.71	0.68	0.52	0.88			
X8	0.82	0.29	0.85	0.82	0.55	0.83	0.92		
X9	0.14	0.07	0.18	0.15	0.55	0.60	0.50	0.46	
X10	-0.31	-0.07	-0.26	-0.28	0.58	0.56	0.33	0.16	0.48

Fig. 2. Correlation matrix for predictor variables

According to Figure 2, there is a strong linear correlation among several predictor variables. This will cause multicollinearity. Handle multicollinearity. Dimension reduction or variable reduction using the PCA method will be used.

TABLE 2
DEACRIPTIVE STATISTICS

Variable	Mean	Standard Deviation	Median
X1	14.127	3.524	13.370
X2	19.290	4.301	18.840
X3	91.970	24.300	86.240
X4	654.900	351.900	551.100
X5	0.096	0.014	0.095
X6	0.104	0.052	0.092
X7	0.088	0.079	0.061
X8	0.048	0.038	0.033
X9	0.181	0.027	0.179
X10	0.062	0.007	0.061

Since the range of data between variables causes unequal values between variables, it is necessary to do feature scaling or standardization of the variables. After performing the feature scaling, the result obtained from PCA is as follows.

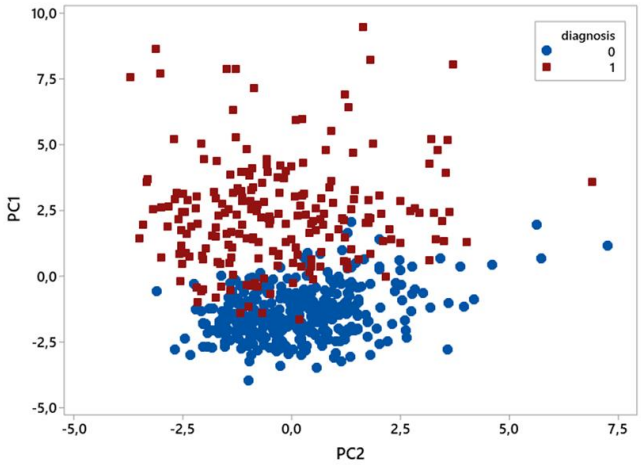

Fig. 3. Scatter Plot of Linear PCA

Figure 3 shows that the components PC1 and PC2, which is a linear combination of predictor variables, in classifying tumor diagnoses between malignant and benign patients,

have not been completely separated. For this reason, dimension reduction will be carried out using the Kernel PCA where they will be brought to a very high (nonlinear) dimension. The kernel function used for the Kernel PCA here is a radial basis function (Gaussian kernel) with optimized sigma parameters.

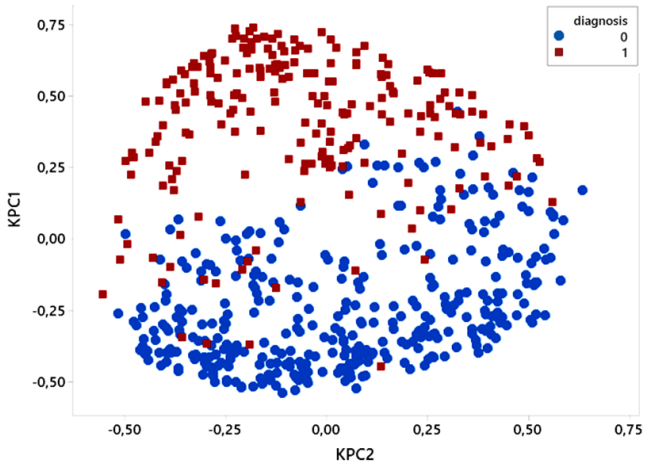

Fig. 4. Scatter Plot of Kernel PCA

Based on Figure 4 using the kernel PCA, the principal component formed can classify patients with benign and malignant tumors and is not mixed as was done by the previous linear PCA. Furthermore, to perform the Hotelling control chart, the principal components of the Kernel PCA are used.

B. *Hotelling's T^2 Control Chart*

Hotelling's T^2 assuming the data follow multivariate normal distribution. The multivariate normal distribution testing of the data is shown in Table 3.

TABLE 3
DISTRIBUTION TESTING

Data	P-Value
PCA	0.000
Kernel PCA	0.000

Based on the p-value, it shows that the data do not follow a multivariate normal distribution. If the data do not follow the multivariate normal distribution, then the control limit of Hotelling's T^2 can be calculate using bootstrap approach.

Using the principal component of the PCA and the kernel PCA, the optimal alpha is needed to get the best results of Hotelling's T^2.

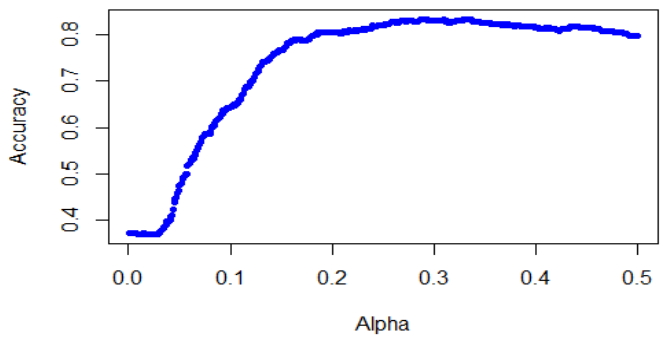

Fig. 5. Hotelling's T^2 accuracy with Linear PCA

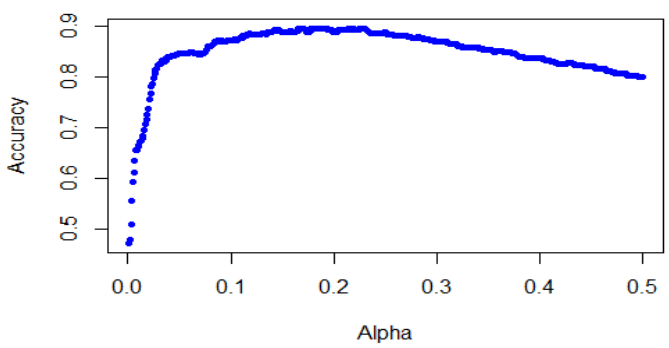

Fig. 6. Hotelling's T^2 accuracy with Kernel PCA

Figure 6 shows that the maximum accuracy is obtained with an alpha that is shown in Table 4.

TABLE 4
OPTIMUM ALPHA

Data	Alpha
PCA	0.331
Kernel PCA	0.192

Table 4 shows that the alpha optimum of Linear PCA is 0.331 and the kernel PCA is 0.192. That alpha will be used for the formation of Hotelling's T^2 for both Linear and Kernel PCA. The Hotelling's T^2 control chart of breast cancer data using linear PCA and Kernel PCA is shown in Figure 7.

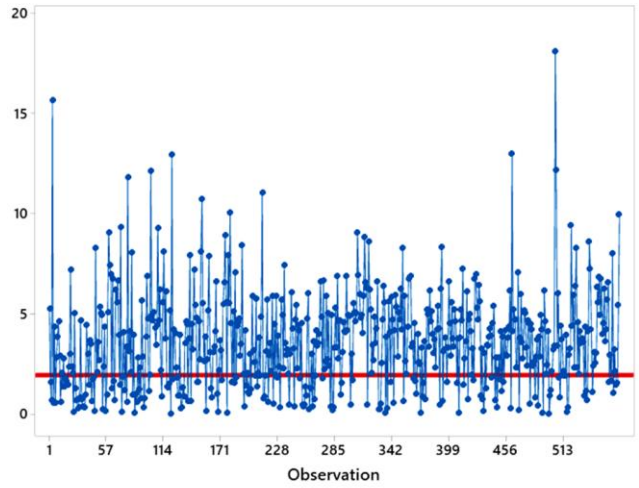

Fig. 7. Hotelling's T^2 control chart with Linear PCA

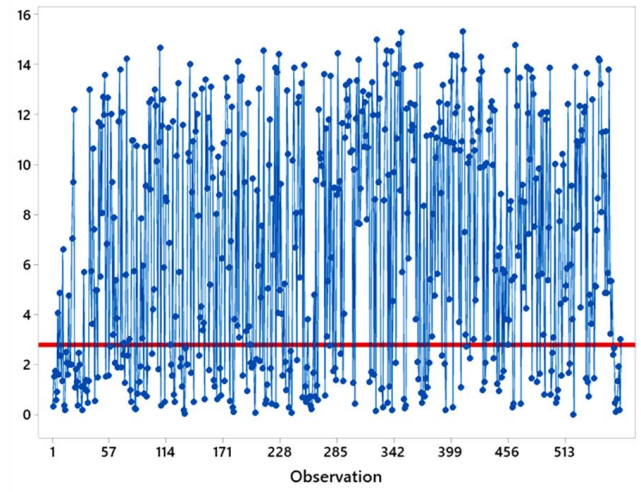

Fig. 8. Hotelling's T^2 control chart with Kernel PCA

According to Figure 8 the Hotelling's T^2 capable to classify patients with benign and malignant breast cancer. The out-of-control point predicts patients with benign breast cancer and the in-control point predicts patients with malignant breast cancer. Suitability of breast cancer classification using Hotelling's T^2 shown in Table 5.

TABLE 5
CONTROL CHART PERFORMANCE

Variables	Accuracy	Sensi	Speci	F1 Score	Balance Accuracy
Linear PCA	0.8348	0.9299	0.675	0.8759	0.8023
Kernel PCA	0.8963	0.9468	0.811	0.9197	0.8790

Table 5 shows the Hotelling's T^2 with Kernel PCA has better performance than the Hotelling's T^2 with linear PCA, since the Hotelling's T^2 with Kernel PCA has higher accuracy, sensitivity, specificity, F1 score and balance accuracy than the Hotelling's T^2 with linear PCA.

C. *Logistic Regression*

Logistic regression is a type of regression analysis in statistics used for the prediction of the outcome of a categorical dependent variable from a set of predictor or independent variables. In this investigation, logistic regression was used as a comparison for the proposed method. The The logistic regression performance with linear PCA is shown in the confusion matrix in Table 6.

TABLE 6
LOGISTIC REGRESSION PERFORMANCE WITH LINEAR PCA

	Benign (0)	Malignant (1)	Precision
Benign (0)	343	14	0.92
Malignant (1)	28	184	0.93

The logistic regression model formed by dimension reduction by linear PCA can predict patients with a diagnosis of benign by 92% and a diagnosis of malignant by 93%.

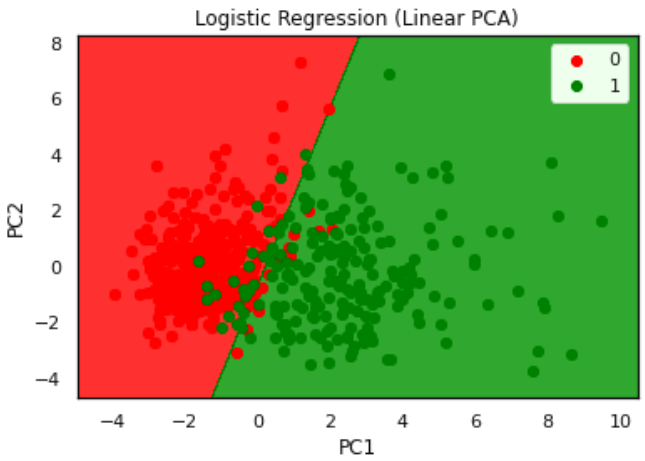

Fig. 9. Logistic regression with linear PCA

Figure 9 shows the grouping results using the logistic regression model with the PC generated. There is a misclassification for benign and malignant diagnoses. The precision of the logistic regression model using the formed PCA is 92.62%. Logistic regression performance with kernel PCA is shown in the confusion matrix in Table 7.

TABLE 7
LOGISTIC REGRESSION PERFORMANCE WITH KERNEL PCA

	Benign (0)	Malignant (1)	Precision
Benign (0)	339	18	0.92
Malignant (1)	30	182	0.91

The logistic regression model formed by dimension reduction by linear PCA can predict patients with a diagnosis of Benign by 92% and a diagnosis of malignant by 91%.

Figure 10 shows the grouping results using the logistic regression model with the PC kernel PC. There is a misclassification for benign and malignant diagnoses. The precision of the logistic regression model using the formed PCA is 91.56%.

D. *Comparison of Hotelling's T^2 with Logistic Regression*

The performance of the Hotelling's T^2 control chart and logistic regression is shown in Table 8.

TABLE 8
METHOD COMPARISON

Classifier	Accuracy
Hotelling's T^2	0.8963
Logistic Regression	0.9156

Based on Table 8, the performance of logistic regression is better in classifying breast cancer diagnoses compared to Hotelling's T^2. These results make sense because the function of logistic regression is to classify. Whereas in Hotelling's T^2, we use the concept of in-control and out of control. However, in predicting the diagnosis of breast cancer, the performance of Hotelling's T^2 with an accuracy value close to 90% can be said to be good.

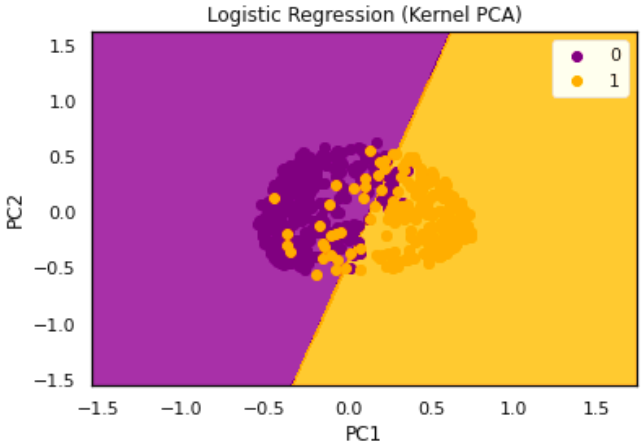

Fig. 10. Logistic regression with Kernel PCA

V. CONCLUSIONS

Hotelling's T^2 control chart using Kernel PCA performed better than PCA to classify benign and malignant breast cancer. The accuracy of Hotelling's T^2 using Kernel PCA is 89.63%. Logistic regression performance is better at classifying breast cancer diagnoses compared to Hotelling's

T^2 since it has a bigger accuracy. These results make sense because the function of logistic regression is to classify. Whereas in Hotelling's T^2. We use the concept of in-control and out of control. However, in predicting the diagnosis of breast cancer. The performance of Hotelling's T^2, with an accuracy value close to 90%, can be said to be good. For future research, it is recommended to use the Multivariate Exponentially Weighted Moving Average (MEWMA) [16].

REFERENCES

[1] D. C. Montgomery, *Introduction to statistical quality control*. John Wiley & Sons, 2020.

[2] M. O. A. Abu-Shawiesh, B. M. Golam Kibria, and F. George, "A robust bivariate control chart alternative to the hotelling's T 2 control chart," *Qual. Reliab. Eng. Int.*, vol. 30, no. 1, pp. 25–35, 2014, doi: 10.1002/qre.1474.

[3] S. Das *et al.*, "Identification of Hot and Cold spots in genome of Mycobacterium tuberculosis using Shewhart Control Charts," *Sci. Rep.*, vol. 2, no. 1, p. 297, 2012.

[4] M. Ahsan, M. Mashuri, and H. Khusna, "Intrusion Detection System Using Bootstrap Resampling Approach Of T2 Control Chart Based On Successive Difference Covariance Matrix," *J. Theor. Appl. Inf. Technol.*, vol. 96, no. 8, pp. 2128–2138, 2018.

[5] P. A. Rogerson and I. Yamada, "Monitoring change in spatial patterns of disease: comparing univariate and multivariate cumulative sum approaches," *Stat. Med.*, vol. 23, no. 14, pp. 2195–2214, 2004.

[6] P. Phaladiganon, S. B. Kim, V. C. P. Chen, J. G. Baek, and S. K. Park, "Bootstrap-based T2 multivariate control charts," *Commun. Stat. Simul. Comput.*, vol. 40, no. 5, pp. 645–662, 2011, doi: 10.1080/03610918.2010.549989.

[7] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley Interdiscip. Rev. Comput. Stat.*, vol. 2, no. 4, pp. 433–459, Jul. 2010, doi: 10.1002/wics.101.

[8] P. Phaladiganon, S. B. Kim, V. C. P. Chen, and W. Jiang, "Principal component analysis-based control charts for multivariate nonnormal distributions," *Expert Syst. Appl.*, vol. 40, no. 8, pp. 3044–3054, 2013, doi: 10.1016/j.eswa.2012.12.020.

[9] D. Kim and I.-B. Lee, "Process monitoring based on probabilistic PCA," *Chemom. Intell. Lab. Syst.*, vol. 67, no. 2, pp. 109–123, 2003, doi: 10.1016/S0169-7439(03)00063-7.

[10] A. Aizerman, "Theoretical foundations of the potential function method in pattern recognition learning," *Autom. Remote Control*, vol. 25, pp. 821–837, 1964.

[11] B. Schölkopf, A. Smola, and K.-R. Müller, "Kernel principal component analysis," *Artif. Neural Networks—ICANN'97*, no. 1, pp. 583–588, 1997, doi: 10.1109/IEMBS.2006.260357.

[12] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A Training Algorithm for Optimal Margin Classifiers," *Proc. 5th Annu. ACM Work. Comput. Learn. Theory*, pp. 144–152, 1992, doi: 10.1.1.21.3818.

[13] M. Ahsan, M. Mashuri, H. Khusna, and M. H. Lee, "Multivariate Control Chart Based on Kernel PCA for Monitoring Mixed Variable and Attribute Quality Characteristics," *Symmetry (Basel).*, vol. 12, no. 11, p. 1838, 2020.

[14] S. Bersimis, S. Psarakis, and J. Panaretos, "Multivariate statistical process control charts: An overview," *Quality and Reliability Engineering International*, vol. 23, no. 5. pp. 517–543, 2007, doi: 10.1002/qre.829.

[15] Y.-S. Sun *et al.*, "Risk factors and preventions of breast cancer," *Int. J. Biol. Sci.*, vol. 13, no. 11, p. 1387, 2017.

[16] N. Sulistiawanti, M. Ahsan, and H. Khusna, "Multivariate Exponentially Weighted Moving Average (MEWMA) and Multivariate Exponentially Weighted Moving Variance (MEWMV) Chart Based on Residual XGBoost Regression for Monitoring Water Quality.," *Eng. Lett.*, vol. 31, no. 3. pp. 1001-1008, 2023.