dx.doi.org/10.24254/CNIB.17.3

# *K*-means y Máquinas de Soporte Vectorial para el Análisis y Clasificación de Señales de Espectroscopía de Inducción Magnética asociadas a Cáncer de Mama.

J.M. Hernández Carrizosa[1], V. Ponomaryov[1], C. A. González Díaz[2*], A. Corzo[3]
[1] Instituto Politécnico Nacional ESIME Culhuacan Cd. México.
[2] Instituto Politécnico Nacional ESM Cd. México.
[3]Universidad del Ejército y Fuerza Aérea/Centro Militar de Ciencias de la Salud, Esc.Mil.Gds.Snd., Cd. México.
*líder de Proyecto: gonzalezantonio@hotmail.com*

*Resumen*— **En la actualidad el sector salud en el primer nivel de atención médica no cuenta con equipos portátiles que permitan un diagnóstico oportuno para detectar cáncer de mama (CaMa). La Escuela Militar de Graduados de Sanidad (EMGS), diseñó un prototipo portátil que utiliza la técnica de Espectroscopia de Inducción Magnética (EIM) como alternativa para detectar neoplasias en tejido mamario de manera no invasiva. El objetivo de este trabajo es encontrar conjuntos específicos de datos de Corrimiento de Fase Inductivo (CFI) y de Ganancia de Corriente Inducida (GCI), obtenidos por este dispositivo, que permitan diferenciar condiciones tisulares asociadas con el CaMa. Utilizamos el algoritmo *k*-means para dividir los espectros correspondientes a tejido sano o con tejido neoplásico benigno. La reducción de dimensionalidad de los espectros se realizó con el Análisis de Componentes Principales (APC) y para la clasificación utilizamos Máquinas de Soporte Vectorial (MSV). Se presentan los resultados obtenidos al clasificar el conjunto de datos que presentó mayor separabilidad entre clases (sanos y CaMa), se alcanzó en promedio una sensibilidad superior al 74% y especificidad mayor al 85% lo que sugiere la viabilidad técnica de la EIM para diferenciar el cambio en la conductividad eléctrica del tejido cancerígeno.**

*Palabras clave*— **Análisis de Componentes Principales, Cáncer de mama, Espectroscopia de Inducción Magnética, *k*-means, Máquinas de Soporte Vectorial.**

## I. Introducción

El Cáncer de Mama (CaMa) es el tipo de cáncer más común a nivel mundial y es la principal causa de muerte en mujeres debido a una neoplasia en países en vías de desarrollo como México [1]. La Espectroscopia de Inducción Magnética (EIM), es un procedimiento no invasivo que permite medir sobre la superficie tisular los cambios de impedancia a diferentes frecuencias, el cambio de conductividad presenta variaciones en relación con las características eléctricas de los diferentes tejidos. En estudios anteriores, se propuso la utilización de mediciones de bioimpedancia a través de campos magnéticos aplicados a múltiples frecuencias como una técnica alternativa para la detección de cáncer de mama de manera no invasiva.

La Escuela Militar de Graduados de Sanidad (EMGS), diseñó un prototipo portátil que utiliza la técnica de Espectroscopia de Inducción Magnética (EIM), como alternativa para detectar neoplasias en tejido mamario de manera no invasiva. Los resultados preliminares de su evaluación clínica indicaron que existen diferencias estadísticamente significativas en frecuencias altas en el rango de los 10 MHz. Se determinó que el incremento en el corrimiento de fase inductivo debido al fenómeno conocido como dispersión *Beta* tiene el potencial para detectar condiciones patológicas del tejido mamario asociadas con el cáncer [2] y [3].

Estudios recientes en el campo de aprendizaje de máquina han planteado diferentes métodos para la clasificación de señales biomédicas, M.R Daliri [4]. Propuso la clasificación para detección de cáncer usando Máquinas de Soporte Vectorial (MSV) y Máquinas de Aprendizaje Extremo (MAE), J. Estrela da Silva, J. P. Marques de Sa y J. Jossinet [5]. Proponen la clasificación de tejido mamario mediante Espectroscopia de Impedancia Eléctrica (EIE) utilizando Análisis Discriminante Lineal (ADL). El objetivo de este trabajo es encontrar conjuntos específicos de datos de Corrimiento de Fase Inductivo (CFI) y de Ganancia de Corriente Inducida (GCI), obtenidas por el dispositivo prototipo, que permitan diferenciar condiciones tisulares asociadas con el CaMa utilizando algoritmos de aprendizaje de máquina como lo son *k*-means, Máquinas de Soporte Vectorial (MSV) y el Análisis de Componentes Principales (ACP) para la reducción de dimensionalidad.

## II. Metodología

1) *Diseño Experimental*: el espectrómetro inductivo portátil mide el CFI y la GCI en cada glándula mamaria, en 136 frecuencias logarítmicamente espaciadas en el ancho de banda de 0.001-100 MHz. Estos datos reflejan las perturbaciones que experimenta un campo magnético en relación con la conductividad del tejido en que se induce.

Conformamos una base de datos de CFI y GCI, seleccionamos de manera aleatoria 895 pacientes femeninas quienes participaron de manera voluntaria en el experimento, previa autorización de consentimiento informado. Las mediciones obtenidas a través del espectrómetro inductivo de las pacientes que participaron en el estudio se dividieron de acuerdo al resultado histopatológico en dos grupos: de Control y CaMa, como se describe en la Tabla 1.

TABLA 1.
GRUPOS EXPERIMENTALES

| Total de pacientes | Mediciones de CFI (2 por paciente) | | Mediciones de GCI (2 por paciente) | |
|---|---|---|---|---|
| 895 | 1,790 | | 1,790 | |
| | Control | CaMa | Control | CaMa |
| | 1,665 | 125 | 1,665 | 125 |

De acuerdo con la información mostrada en la Tabla 1, el grupo de Control corresponde al 93% del total de datos en contraste con el 7% restante (CaMa).

Con el fin de evidenciar las características más importantes de los espectros y minimizar o neutralizar aquellas mediciones que no contribuyen al análisis experimental se utilizó como vector de referencia la media muestral correspondiente a 30 mediciones sin la presencia de tejido tanto de CFI como de GCI.

2) *Estandarización*: se utilizó este método inicial para eliminar frecuencias en donde las mediciones de inducción magnética permanecen constantes con o sin tejido mamario. En (1) se define matemáticamente este proceso. Obtuvimos el valor absoluto de la diferencia entre el vector de referencia y los espectros de CFI y GCI. En la Fig.1, se observan las gráficas de los espectros estandarizados.

$$x_E^{(n)} = \left| v_R - x^{(n)} \right| \tag{1}$$

$x_E^{(n)} = Espectro\ estandarizado\ \{n|n = 1,2,3, \dots,1665\}.$
$x^{(n)} = Espectro\ \ original\ \{n|n = 1,2,3, \dots,1665\}.$
$v_R = Vector\ de\ referencia.$

3) *Eliminación de mediciones no significativas*: para el caso de los espectros estandarizados de CFI eliminamos 32 de 136 frecuencias quedando como resultado final 104 frecuencias en donde el valor absoluto fue distinto de cero. Para el caso de los espectros de GCI eliminamos bajo el mismo criterio 48 frecuencias por lo que el resultado final en este caso fue de 88 frecuencias de medición.

4) *Agrupamiento de espectros mediante el algoritmo k-means*: el algoritmo *k*-means [6]. Es un método que dividió automáticamente a los grupos de control en 3 regiones ($k$=3). Dado el conjunto X donde $\left\{x_E^{(1)}, x_E^{(2)}, x_E^{(3)}, \dots, x_E^{(n)}\right\}$, representan los espectros estandarizados del grupo de control, es dividido en subgrupos "clusters", mediante un procedimiento iterativo que comienza seleccionando un centroide de manera aleatoria asignándole los datos más cercanos, en (2) se indica este procedimiento.

$$c^{(i)} \coloneqq j\ que\ minimiza\ \left\| x_E^{(n)} - \mu_j \right\| \tag{2}$$

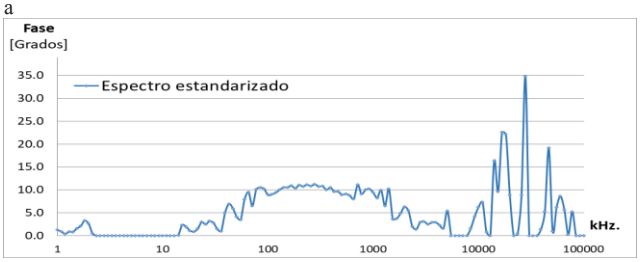
$c^{(i)} = índice\ del\ centroide\ mas\ cercano\ a\ x^{(n)}.$
$x_E^{(n)} = Espectro\ estandarizado\ \{n|n = 1,2,3, \dots,1665\}.$
$\mu_j = valor\ del\ j'\ centroide.$

Con base en la ecuación anterior se realizaron 100 iteraciones recalculando los centroides en cada ciclo basándose en el conjunto de datos previamente agrupado, dada la asignación de cada elemento $x_E^{(n)}$ a un centroide, el algoritmo redefine para cada centroide, la media de los puntos que fueron asignados a éste como se indica en (3).

$$\mu_k \coloneqq \frac{1}{|C_k|} \sum_{i \in C_k} x_E^{(n)} \tag{3}$$

$\mu_k = media\ de\ los\ puntos\ asignados\ al\ centroide\ k.$
$C_k = grupo\ de\ elementos\ asignados\ al\ centroide\ k.$

a




b


Fig. 1. Espectros estandarizados, (a) CFI, (b) GCI.

5) *Clasificación de espectros de CFI y GCI utilizando ACP y MSV*: se seleccionaron los subgrupos de CFI y GCI en donde existió mayor diferencia estadística respecto al grupo de CaMa, los espectros fueron procesados a través del ACP antes de ser sometidos al clasificador. El ACP es una técnica matemática de reducción de dimensionalidad, bajo esta técnica se insertaron en un subespacio lineal de menor dimensionalidad las mediciones de CFI (correspondientes a 104 frecuencias) y de GCI (correspondientes a 88 frecuencias), dicho subespacio describió tanto como fue posible la varianza de los datos, esto se logró al encontrar una base lineal en donde la cantidad de varianza fue la máxima [7].

Se diseñó un clasificador mediante MSV utilizando un conjunto para entrenamiento y uno de prueba. En la etapa de entrenamiento se generó un hiperplano que dividió el grupo de datos positivos (CaMa), del conjunto de datos negativos (sanos), este algoritmo seleccionó el límite máximo en el espacio de características entre los dos conjuntos [8]. Posteriormente se evaluó el clasificador con el conjunto de prueba. Para entrenar el clasificador se utilizó MSV con kernel de función de base radial Gaussiana, cuya expresión matemática se define en (4).

$$K(x_i, x_j) = \exp\left(-\gamma \left\| x_i - x_j \right\|^2\right) \tag{4}$$

$x = ejemplo\ de\ entrenamiento$
$\gamma = \frac{1}{2\sigma^2}\ donde\ \sigma^2\ es\ la\ varianza$
$\left\| x_i - x_j \right\|^2 = distancia\ euclidea\ al\ cuadrado$

6) *Criterios de medición de desempeño*: se seleccionó de manera aleatoria el 50 y 40% del total de datos para la etapa de prueba y se realizó cada experimento 100 veces. La exactitud se calcula considerando la proporción entre las muestras verdaderamente clasificadas y el número total de muestras, la sensibilidad es la proporción entre los casos detectados como positivos y el total de casos positivos y la

especificidad es la proporción entre los casos detectados como negativos y el total de casos verdaderamente negativos.

## III. RESULTADOS

Los resultados obtenidos al aplicar el algoritmo *k*-means, al grupo de control se muestran en la Fig. 2, se realizaron 30 iteraciones a fin de confirmar que los datos se agruparan de la misma forma y de manera constante debido a que el algoritmo selecciona aleatoriamente el centroide inicial.

En la Fig. 3, se muestran gráficamente el vector promedio de un conjunto específico de datos, definido como "subgrupo 3", el cual se discrimina fácilmente respecto al vector promedio del grupo de CaMa, siendo más evidente en altas frecuencias. De acuerdo con lo observado en la Fig. 3, se realizó la prueba t-student en las frecuencias de campo magnético en donde se encontró mayor distancia entre los grupos, las Tablas 2 y 3 muestran diferencias estadísticamente significativas en estas frecuencias.

Respecto a la clasificación realizada mediante MSV, Las gráficas de las Figs. 4 y 5, muestran la sensibilidad y especificidad alcanzadas por el clasificador de espectros de CFI, se consideró mostrar únicamente los resultados obtenidos utilizando el 40% de los datos para prueba ya que se observó menor desviación estándar. Se realizaron pruebas experimentales con 30 componentes principales, la sensibilidad se mantuvo constante en un promedio del 70%, y la especificidad en un promedio del 90%.

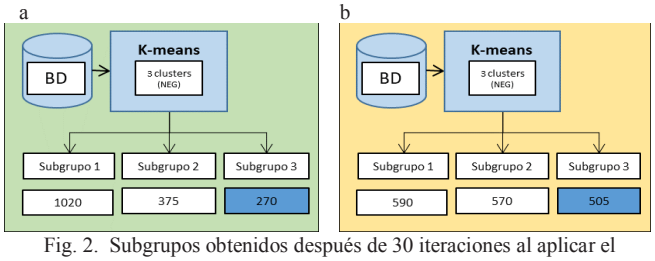



Fig. 2. Subgrupos obtenidos después de 30 iteraciones al aplicar el algoritmo k-means, (a) CFI, (b) GCI.

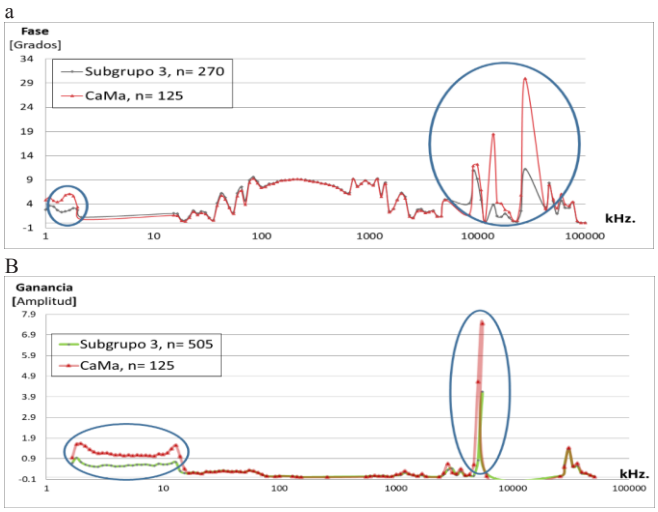



Fig. 3, Comparación de los vectores promedio del subgrupo 3 y el grupo CaMa (a) CFI, (b) GCI.

TABLA 2.
T-STUDENT PARA MUESTRAS INDEPENDIENTES (CFI)

| KHz | Media Subgrupo 3 Control | Media CaMa | t-value | fd | P |
|---|---|---|---|---|---|
| 1 | 2.810 | 4.858 | -5.4586 | 152.84 | 1.90E-07 |
| 1.089 | 3.576 | 5.370 | -4.8029 | 185.74 | 3.21E-06 |
| 1.2915 | 2.733 | 4.399 | -4.8383 | 190.64 | 2.69E-06 |
| 1.4065 | 2.285 | 4.836 | -6.9838 | 156.04 | 7.80E-11 |
| 1.5317 | 2.439 | 5.800 | -8.4154 | 148.78 | 3.00E-14 |
| 1.6681 | 2.678 | 6.091 | -9.3609 | 158.84 | < 2.2e-16 |
| 1.8165 | 3.157 | 5.689 | -9.1696 | 199.27 | < 2.2e-16 |
| 14065.272 | 3.864 | 18.516 | -12.504 | 146.21 | < 2.2e-16 |
| 27825.594 | 11.245 | 30.001 | -21.599 | 196.51 | < 2.2e-16 |

TABLA 3.
T-STUDENT PARA MUESTRAS INDEPENDIENTES (GCI)

| KHz | Media Subgrupo 3 Control | Media CaMa | t-value | fd | P |
|---|---|---|---|---|---|
| 1.817 | 0.941 | 1.610 | -12.858 | 152.87 | < 2.2e-16 |
| 1.978 | 0.732 | 1.639 | -13.401 | 138.66 | < 2.2e-16 |
| 2.154 | 0.613 | 1.519 | -12.823 | 132.69 | < 2.2e-16 |
| 2.346 | 0.551 | 1.357 | -12.004 | 130.38 | < 2.2e-16 |
| 2.555 | 0.515 | 1.241 | -11.639 | 130.06 | < 2.2e-16 |
| 2.783 | 0.509 | 1.170 | -11.601 | 130.65 | < 2.2e-16 |
| 3.030 | 0.569 | 1.183 | -11.456 | 131.59 | < 2.2e-16 |
| 12.915 | 0.731 | 1.563 | -12.693 | 135.99 | < 2.2e-16 |
| 5054.797 | 0.829 | 4.610 | -11.661 | 126.15 | < 2.2e-16 |
| 5504.790 | 4.141 | 7.401 | -14.54 | 158.28 | < 2.2e-16 |

Las gráficas de la Fig. 6 y 7, muestran el comportamiento del clasificador de GCI, utilizando los mismos criterios. Se obtuvieron idénticos porcentajes de sensibilidad y especificidad (70% y 90% respectivamente).

La Tabla 4, muestra los resultados promedio obtenidos por los clasificadores después de 100 iteraciones, en el caso de los espectros de CFI se requieren únicamente 3 componentes principales, lo que reduce el tiempo de procesamiento. Los mejores resultados para los datos de GCI se obtuvieron con 24 componentes principales.

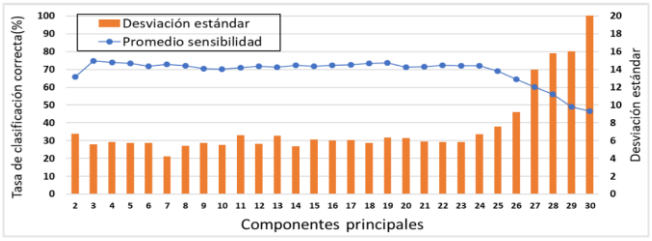



Fig. 4, Sensibilidad del clasificador de CFI y 40% de datos para prueba.

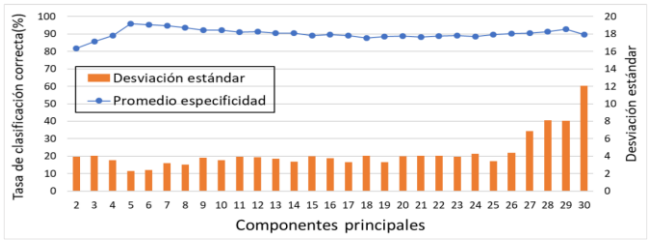



Fig. 5, Especificidad del clasificador de CFI y 40% de datos para prueba.

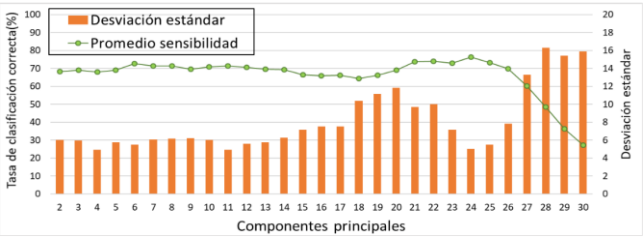



Fig. 6. Sensibilidad del clasificador de GCI y 40% de datos para prueba.

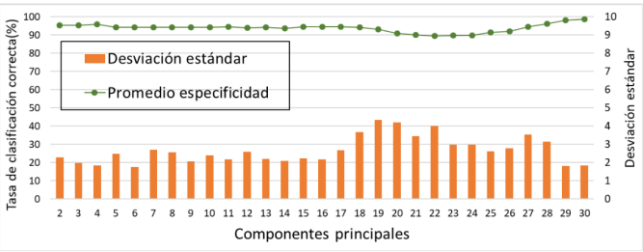



Fig. 7. Especificidad del clasificador de GCI y 40% de datos para prueba.

## IV. Discusión

El análisis y clasificación de los espectros de CFI Y GCI se dividió en 3 etapas: la estandarización, la segmentación del grupo de control y la clasificación de los datos. La estandarización en función de los vectores de referencia permitió identificar frecuencias en donde no se producen cambios en las mediciones ante la presencia de tejido y enfatizó aquellas en donde existió mayor variabilidad. La segmentación del grupo de control mediante *k*-means sirvió para encontrar un conjunto de datos susceptibles de ser clasificados, para el caso del CFI correspondió al 16% del total de datos, mientras que para la GCI fue del 30%. Obtener la mejor clasificación de CFI en función de 3 componentes principales implica que es posible reducir a 3 características la información contenida en 104 mediciones, las gráficas mostradas en las Figs. 4 y 5 indican que no hay un cambio sustancial en la clasificación al incrementar el número de componentes, esto se debe a que en una región tridimensional se puede definir la máxima varianza entre las 2 clases, una condición distinta se observa en el caso de la GCI, a pesar de que la sensibilidad se mantiene en un promedio del 70%, la mejor clasificación y menor desviación estándar se obtuvieron con 24 componentes principales como se observa en la Fig. 6, la desviación estándar en este caso fue la que definió la selección de componentes principales, esto no produjo un cambio importante en la especificidad como se muestra en la Fig. 7. Los resultados de las Tablas 2 y 3 indican que las mediciones de inducción magnética tienden a diferenciarse entre el grupo de Control y CaMa a bajas frecuencias (alrededor de los 2 kHz.) y a altas frecuencias (5, 14 y 27 MHz.), debido probablemente a los fenómenos conocidos como dispersiones dieléctricas *Alpha* y *Beta*. Se corroboró lo expuesto en [3], al encontrar diferencias estadísticamente significativas en el CFI en la frecuencia de 14.06 MHz. La clasificación mediante MSV permitió obtener una sensibilidad promedio del 74% y especificidad del 85%.

TABLA 4.
MEJORES RESULTADOS DE CLASIFICACIÓN

| | | CFI | | |
|---|---|---|---|---|
| Método | Exactitud (%) | Sensibilidad (%) | Especificidad (%) | CP |
| PCA | **82.2±3** | **74.7 ± 5** | **85.7 ± 4** | 3 |
| | | GCI | | |
| PCA | **87.2±2** | **76.2 ± 5** | **89.9± 3** | 24 |

## V. Conclusión

Al combinar diversas técnicas de aprendizaje de máquina para el análisis y clasificación de los espectros de CFI y GCI se comprobó que la EIM tiene viabilidad técnica para detectar condiciones patológicas en la glándula mamaria asociadas con el cáncer de mama. Los datos que permitieron alcanzar una sensibilidad del 74% constituye una base fundamental para observar el comportamiento de la inducción magnética en este órgano, sin embargo, se requiere de un estudio riguroso de los conjuntos de datos ya obtenidos que sirva para extraer y seleccionar características específicas que permitan generalizar los resultados obtenidos en este trabajo.

## Reconocimientos

Los autores agradecen al Instituto Politécnico Nacional, al CONACYT (proyecto 220347) por el apoyo brindado y a la EMGS por permitirnos trabajar con el material que sustenta la patente: “Sistema Inductor-Sensor para detección de cáncer en glándula mamaria a través de campos magnéticos (patente: 323903 septiembre de 2014)”.

## Bibliografía

[1] L. A. Torre, F. Bray, R. L. Siegel, J. Ferlay, J. Lortet-Tieulent, A. Jemal, “Global cancer statistics”, 2012. CA: A Cancer Journal for Clinicians. 65(2):87–108, 2015.

[2] C. A. González, L. M. Lozano, M. C. Uscanga, J. G. Silva and S. M. Polo, “Theoretical and Experimental Estimations of Volumetric Inductive Phase Shift in Breast Cancer Tissue”, Journal of Physics: Conference Series 434, 2013.

[3] C. A. González, M. C. Uscanga, L. M. Lozano, J. L. Ortíz, J. A. González, and C. I. Guerrero-Robles, “Clinical Evaluation of Inductive Spectrometer to detect Breast Cancer.” VII Latin American Congress on Biomedical Engineering CLAIB 2016, Bucaramanga, Santander, Colombia, October 26th -28th, 2016 IFMBE Proceedings, 678-681.

[4] M. R Daliri, “Combining extreme learning machines using support vector machines for breast tissue classification”, Comput Methods Biomech Biomed Eng 2015; 18: 185 – 191.

[5] J. Estrela da Silva, J.P. Marques de Sa, J. Jossinet, “Classification of breast tissue by electrical impedance spectroscopy”, Med Biol Eng Comput. 38:26–30, 2000.

[6] A. Coates, H. Lee, A.Y. Ng, “An analysis of single-layer networks in unsupervised feature learning.”, In: 14th International Conference on AI and Statistics. pp. 215– 223, 2011.

[7] L. van der Maaten, E. Postma and J. van den Herik, “Dimensionality Reduction: A Comparative Review”, Technical Report TiCC-TR 2009-005, Tilburg University, 2009.

[8] Z. Nematzadeh, R. Ibrahim and A. Selamat, “Comparative studies on breast cancer classifications with k-fold cross validations using machine learning techniques.”, 10th Asian Control Conference (ASCC). 2015.