Investigación CIENTÍFICA
Vol. 14, No. 2 (2020)
Jornadas de Investigación UAZ – 2020

Predicción de tumores benignos o malignos a través de un modelo de selección de características

María de la Luz Escobar, José María Celaya Padilla, y José Ismael de la Rosa

Universidad Autónoma de Zacatecas,
Unidad Académica de Ingeniería Eléctrica
Avenida López Velarde 801, Zacatecas, Zac., CP 98000

escobarmaria50@uaz.edu.mx

Resumen: El cáncer de mama es una de las principales causas de letalidad entre la población femenina. El análisis de características en una mamografía ayuda en parte a predecir el riesgo de cáncer de mama. Por ejemplo, la densidad de masas (o lesiones) se correlaciona con los tumores, para esto, la segmentación de tumores en imágenes médicas proporciona datos que se pueden obtener, y utilizar como indicador para la clasificación de lesiones (benignas / malignas). En este trabajo propone, un modelo de clasificación de lesiones (tumores), basado en la extracción de características a través de la región de interés (ROI), en conjunto con un modelo de regresión múltiple para la clasificación. La Textura, forma, color y descriptores estadísticos de características se utilizan para obtener información de imágenes. Estas características se evalúan para reducir las características utilizando un modelo de selección de características. Posteriormente se realiza una evaluación en términos del área bajo la curva (AUC), cuyo resultado proporcionó un máximo de 0.9263 para la clasificación de tumores.

***Palabras clave*:** Cáncer de mama, Selección de Características, Curvas ROC.

Abstract*:* Breast cancer is one of the leading causes of death in women. The features that are taken from mammograms can help in partly to predict the risk of breast cancer. For example, the density of masses (lesions) is correlated with tumors, for this reason, the segmentation of tumors on medical images provide data that can be obtained and used as an indicator for the classification of lesions (benign/malignant). In this work, a classification model of lesions (tumors) is proposed, based on the extraction of features of the region of interest (ROI), using a multiple regression model to classify them. Texture, shape, color and statistical feature descriptors are used to obtain image information. These features are evaluated to reduce the model using feature selection techniques. After the selection of features, an evaluation was made in terms of the area under the curve (AUC), obtaining a maximum of 0.9263 for the classification of tumors.

Keywords: Breast cancer, Selection Features, Curves ROC.

1. Introducción

El nombre de cáncer refiere a un conjunto de enfermedades que se presentan en el cuerpo. El cáncer de mama ocupa el segundo lugar en decesos entre la población femenina a nivel mundial, solo superado por el cáncer de pulmón [1, 2]. Las estimaciones del año de 2019 presentan 2 088 849 millones de nuevos casos y 629 679 millones de defunciones [1]. En México se diagnosticaron 27 283 y se presentaron 6 884 muertes seguidas por el cáncer cervicouterino [2]. Sin embargo, algunas investigaciones refieren, a que mediante un diagnóstico temprano existe la posibilidad de un descenso en la letalidad femenina, ya que ésta sería más vulnerable a un tratamiento curativo, aumentando la esperanza de vida [3 - 5].

2. Marco Teórico

El cáncer de mama es un tipo de cáncer mortal entre las mujeres en todo el mundo, alcanzando 9,6 millones de muertes y 2 millones de casos en 2019 [1-2] y cantidades similares para 2019 [3]. Las Mamografías se utilizan como una herramienta para el diagnóstico de cáncer de mama, y cuando se encuentra en sus primeras etapas (diagnóstico precoz), los tratamientos son más eficientes, contribuyendo a reducir el número de muertes por esta enfermedad [4 - 6]. El análisis asistido por computadora permite ayudar a visualizar la forma, el contorno, la densidad y el perímetro de la masa, y cuya observación permite realizar una estimación de la clasificación de la lesión de cáncer de mama [7 - 12]. Las diferentes propuestas en la literatura se sustentan en el desarrollo de técnicas informáticas para el mejoramiento de características para un diagnóstico óptimo de la enfermedad. Por ejemplo, Galván *et al.* [5] proponen un modelo multivariado para la clasificación de lesión de tumores benignos o malignos, mediante un algoritmo genético realizando un análisis de características morfológicas de las lesiones y obteniendo una clasificación. Otros estudios proponen modelos cuyo foco principal es la reducción de falsos positivos, utilizando técnicas de optimización en imágenes para clasificar las lesiones [13, 14]. Por otro lado, Hernández *et al* [15], proponen una reducción de falsos positivos mediante la clasificación grasa y grasa glandular

mediante un conjunto de características determinadas por micro-clasificaciones.

Otros enfoques de análisis multivalente han demostrado que la información de un pronóstico y factores predictivos se puede obtener al identificar el cáncer de mama en sus primeras etapas [16].

Entre las diferentes técnicas de procesamiento de imágenes digitales y reconocimiento de patrones que se han aplicado en la literatura para la detección de cáncer de mama, se encuentra el uso de información mutua y una selectividad para el diagnóstico, utilizada cuando la información está uniformemente distribuida [17]. Por tanto, este artículo propone una metodología novedosa para análisis de características obtenidas dentro de una imagen mamografía, utilizando un enfoque de selección de características y un modelo de regresión multivalente como clasificador de los tumores benignos o malignos. El objetivo principal de la presente propuesta es proporcionar un modelo multivariado, capaz de clasificar los tumores que se someten a una modelo computacional para confirmar tal diagnóstico. El modelo propuesto tiene la ventaja de reducir el número de funciones, simplificando así la implementación futura, permitiendo el diagnóstico asistido de tumores en etapa temprana.

El artículo se organizado en cinco partes. La primera comienza con una breve introducción sobre el cáncer de mama. A continuación, se presenta una visión general del estado del arte. Posteriormente se describen el materiales y métodos implementados en la metodología. Finalmente, se presentan las conclusiones y las referencias.

3. Materiales y Métodos

En esta sección se presenta con detalle el análisis de la implementación del modelo propuesto para la clasificación y selección de características en imágenes mastográficas (Fig. 1). Primeramente, un radiólogo delimita una región de interés (ROI) sobre el cual se evalúa el tumor sospechoso. El ROI es utilizado por un proceso automatizado que extraer un conjunto características, y cuya información generada define un modelo regresión lineal. Al término del proceso anterior, se llevó a cabo la especificación del modelo y la estimación de parámetros. El estimador de mínimos cuadrados es utilizado como criterio de minimización en el modelo de regresión lineal sugerido, y mediante el cálculo del área bajo la curva se realiza la clasificación de tejidos sanos o cancerosos. A través de un estudio de simulación, se lleva a cabo la optimización de variables del modelo de regresión lineal, y mediante la implementación de tres métodos de selección de variables se obtienen nuevos modelos con información relevante. Finalmente, el rendimiento de las técnicas de selección de características se evalúa mediante el error cuadrático medio.

3.1 Selección de características basado en mínimos cuadrados

En este estudio se implementa la técnica de mínimos cuadrados que es usada para explorar el mejor ajuste en un conjunto de puntos de datos. El conjunto de datos está centrando el conjunto de datos X_{ij}. En esta etapa, el vector x_i observado se transforma en vector y_j.. Cálculo de regresión múltiple (Eq. 1).

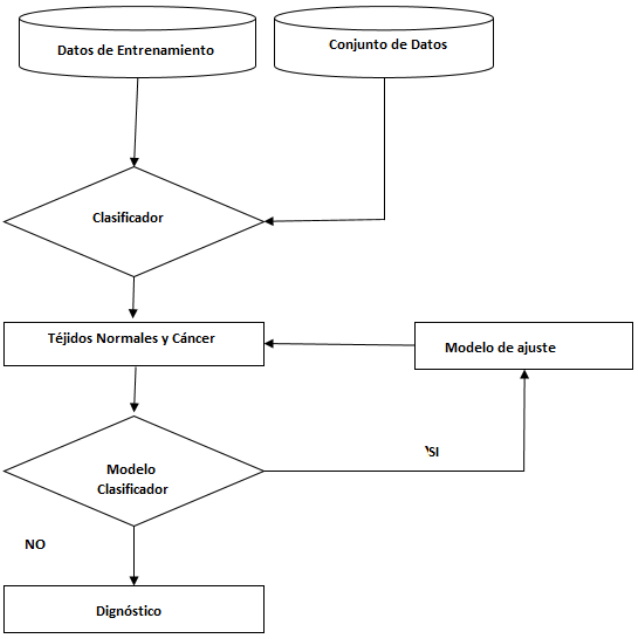

Fig. 1. Diagrama selección de características.

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots \beta_n X_n + \epsilon \qquad (1)$$

donde β_0 representa el origen, $\beta_1 + \beta_2 + \beta_3 + \beta_3 + \ldots + \beta_n$ representa los coeficientes de la pendiente de la recta y $X_1 + X_2 + X_3 + X_3 + \cdots + X_n$ y representan las variables independientes, y_i es la variable dependiente, La ecuación anterior se puede expresar de la siguiente forma Eq. 2.

$$Y = X\beta + \epsilon \qquad (2)$$

donde β es un vector de parámetros desconocidos Y es un vector d observaciones, Y es la variable dependiente de la muestras y X la variable independiente. Y representa un vector de estimaciones de β y ϵ el error.

3.2 Mínimos cuadrados

El análisis de regresión es el método estadístico que permite evaluar una relación entre las variables e identificar cuales tienen relación con los tejidos cancerosos o sanos, de forma que la regresión se obtiene ajustando las observaciones mediante el mínimos cuadrados (Eq. 3).

$$Y = (XX^t)^{-1} X^t \beta \qquad (3)$$

donde X^T es la traspuesta de X, y Y es una aproximación del modelo.

Expresada en logaritmo neperiano se tiene la Eq. 4.

$$Y = e^{\beta i} \quad (4)$$

Si la imagen está en presencia de tejido canceroso AUC >0.5; en caso contrario, es tejido normal.

3.3 Residuo y mínimos cuadrados

Las técnicas de selección de variables predictores para el modelo de regresión lineal sugerido, tienen como finalidad evaluar la que promociona información redundante. Las técnicas paso-paso, selección hacia adelante y selección hacia atrás son utilizadas para disminuir el conjunto de atributos dentro de una muestra de datos. Los algoritmos consisten en generación de todas las combinaciones posibles de un conjunto de variables, de tamaño 1 hasta p, siendo p el número total de variables. El criterio de selección se basa en el erro cuadrático medio (RMSE). Las técnicas se describen a continuación:

La selección hacia atrás comienza con todas las variables del modelo, enseguida se van eliminando todas las variables de acuerdo menos al criterio de selección

La selección hacia adelante comienza con una variable en el modelo, el cual se va agregando características cumplan con el criterio de selección

Selección paso-paso es una combinación de los procedimientos anteriores. En cada paso se introduce una variable independiente que no se encuentre en la ecuación, este método puede eliminar variables que se encuentren en el modelo (Eq. 5)

$$y_{new} = \begin{cases} \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots \beta_n X_n & para\ toda \geq 0.5 \\ 0 & de\ otra\ manera \end{cases} \quad (5)$$

El AIC es el criterio de clasificación utilizado para obtener los nuevos modelos, y cuyo valor superior a 0.5 para la identificación de cáncer, de otra forma se clasificará como tejido sano.

El conjunto de datos utilizado en este trabajo se obtuvo del "Repositorio Digital Portugués de Cáncer de Mama" (BCDRD1 y BCDR-D2) conjunto de datos [18]. De los cuales, se obtuvieron 500 muestras que pertenecen a dos tipos de lesiones (Benignas versus hallazgos malignos), tales lesiones fueron corroboradas por la biopsia. Los datos incluyeron: datos clínicos y características presentes en imágenes.

Las características del conjunto de datos se agrupan en: estadísticas de primer orden, análisis de textura, forma y ubicación del contorno de la lesión identificada por el radiólogo. Un total de 28 características fueron extraídas de la ROI de la imagen. La Tabla 1 muestra la lista completa de características utilizadas en esta investigación. El conjunto de datos se compone de dos conjuntos de datos BCDR-D1 con 455 imágenes y BCDR-D2 con 145 imágenes, haciendo un total de 600 imágenes.

3.4 Modelo de Generación

Para obtener un modelo representativo, el conjunto de datos de la sesión anterior, se extrajeron dos sub conjuntos de datos. El primer subconjunto representa el conjunto de entrenamiento, y el segundo, datos de prueba.

Tabla 1. Características extraídas de las imágenes mastográficas

Estimación del modelo
Media
Desviación estándar
Valor Máximo
Curtosis
Suavidad
Área
Perímetro
Centro de masas
Circularidad
Elongación
Forma
Solido
Extensión
Energía
Contraste
Correlación
Varianza
Homogeneidad
Suma Porcentaje
Suma de Varianzas
Suma de Entropías
Entropía
Diferencias de varianzas
Diferencias de entropía
Correlación1
Correlación2

A continuación, se emplea el análisis regresivo de los sub-conjuntos generados, los cuales son configurados de la siguiente manera:

A partir de la información proporcionada por los píxeles en escalas de grises en la ROI de la imagen mastográfica, la cual fue delimitada por un radiólogo (Tabla 1), es utilizada como entrada al algoritmo mínimos cuadrados. Los datos de entrenamiento y prueba comienzan con 28 características. Las muestra inicial consta de un subconjunto de 150 de datos observados y 455 de entrenamiento (Fig. 2). El trabajo se llevó a cabo usando herramienta de software MATLAB´s.

4. Resultados y Discusión

Primeramente, se implementa el algoritmo de aprendizaje automático cuya finalidad es minimizar la predicción incorrecta.

Los datos se forman aleatoriamente para asegurarse que los conjuntos de entrenamiento y de prueba sean similares. Una vez procesado el modelo de regresión, se utiliza el conjunto de entre-

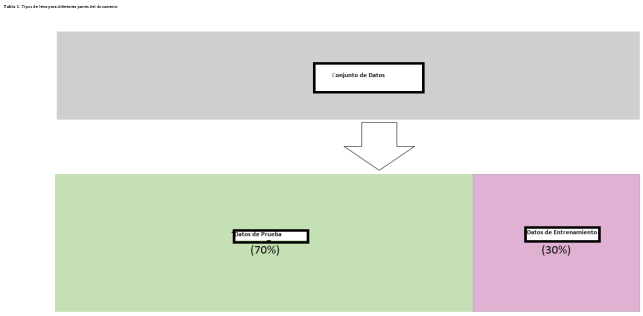

Fig. 2. Diagrama de comparación de valores predictivos.

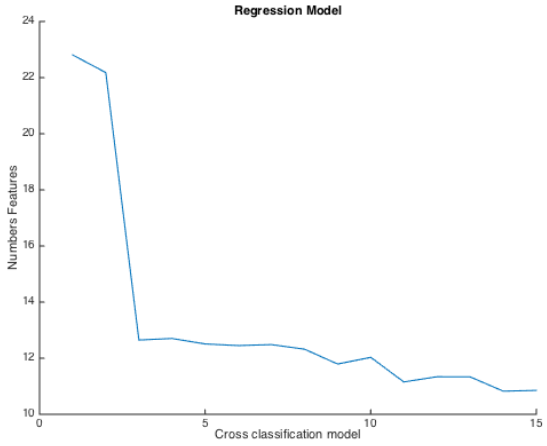

Fig. 3. Diagrama de comparación de valores predictivos.

namiento y pruebas para realizar las comparaciones de los modelos de predicciones.

En la Figura 3 se muestra la gráfica resultante de la estimación del modelo predictor, donde las coordenadas verticales representan el número de características y en las coordenadas horizontales el resultado del modelo multivariado. Como se puede observar en la gráfica los sesgos presentes, se describen que modelo está mal especificado, las estimaciones de los coeficientes pueden resultar considerablemente sesgadas.

La optimización del modelo predictor, parte de la primicia, si sabemos que cuantos más predictores incluyamos en el modelo, mejores predicciones con un menor sesgo, pero a la vez menos precisión en las variables predictora. El propósito de la siguiente etapa es la selección de características como un problema selección y comparación de modelos.

La eliminación de la información redundante, se realiza a partir de técnicas de selección paso-paso, hacia atrás y hacia adelante. Los estudios del análisis multivariante para los datos de prueba y de entrenamiento son mostrados en las Tablas 2 y 3.

El experimento comienza con una muestra de 143 observaciones de entrenamiento con 115 grado de libertad (Tabla

Tabla 2. LINEAR REGRESSION MODEL
Intercept Distribution Binominal

Modelo Estimación	Estimación	SE tsat pValue	tsat	P Value
Media	- 247.07	98.835	-2.4998	0.012428
Desviación Estándar	-92.194	95.535	1.3156	0.18832
Valor Máximo	-5.5383	10.168	-0.54469	0.58597
Curtosis	1.1953	0.48295	2.475l	0.013325
Suavidad	-3.3945	1.8191 l	-1.866	0.062042
Área	5.59E-06	9.49E06	0.58872	0.55605
Perímetro	-0.001419	0.0015044	-0.94363	0.34536
Centro de masas	-2.7872	1.4832	1.8792	0.060216
Circularidad	-11.439	7.4205	-1.5416	0.12317
Elongación	5.4699	3.5694	1.5324	0.12541
Forma	-10.934	477.74	0.022887	0.98174
Solido	6.8514	7.2999	0.93856	0.34796
Extensión	1.1228	6.1756	-0.18181	0.85573
Energía	-63.644	83.822	-0.75927	0.44769
Contraste	-0.67031	0.34019	-1.9704	0.048794
Correlación	7.805	13.809	-0.56519	0.57194
Varianza	0.20673	0.44625	0.46326	0.64318
Homogeneidad	89.578	51.195	1.6665	0.095622
Suma Porcentaje	3.5749	1.6208	2.2057	0.027408
Suma de Varianzas	-0.051902	0.11099	-0.46762	0.64006
Suma de Entropías	-6.5562	19.066	-0.34387	0.73094
Entropía	7.4365	11.542	0.6443	0.51938
Diferencias de varianzas	0	0	NAN	NAN
Diferencias de entropía	30.066	16.847	1.7847	0.074315
Correlación1	2 -30.999	41.794	-0.74172	0.45826
Correlación2	2	-2.2882	16.068	-0.14241

2). La dispersión Chi ^ 2-estadístico frente al modelo constante: 60,7, valor p = 0,000211.

En el siguiente experimento se extrae un subconjunto de 580 observaciones aleatoriamente, 115 grados de libertad. En la Tabla 3 se muestra el rendimiento de dispersión: estadística de Chi ^ 2 frente a modelo constante: 60,7, valor p = 0,000211 y un criterio de AUC = 0.98.

El criterio de AUC sugiere que el modelo paso-paso presenta un mejor rendimiento en comparación con los otros métodos implementados, obteniendo un valor de 0.8558 de positividad de prueba (Fig. 4 y 5).

Finalmente, en la Tabla 4 se muestra los datos estadísticos de cada uno de las técnicas de selección. El resultado final es ahora mucho más interesante; la estadística de Ch^2 muestra un mejor rendimiento en la prueba de selección hacia-atrás, con un nivel de confianza mayor al 50 %, mientras que los otras técnicas presentan rendimiento abajo del 50%.

Tabla 3. LINEAR REGRESSION MODEL
Intercept Distribution Binominal

Modelo Estimación	Estimación	SE tsat pValue	tsat	PValue
Media	-12.704	98.835	-0.28122	0.77854
Desviación estándar	53.366	95.535	1.3156	0.18832
Valor Máximo	17.699	6.3477	1.5062	0.13201
Curtosis	0.29131	0.26561	1.0968	1.0968
Suavidad	-2.9555	1.1884	-1.866	0.012883
Área	3.85E-06	9.49E06	0.64839	0.51673
Perímetro	-0.00081	0.0015044	0.0013	0.54962
Centro de masas	-1.1603	0.86833	-1.3362	0.18147
Circularidad	-10.531	6.3045	-1.6705	0.09483
Elongación	8.4131	2.8192	2.9842	0.0028434
Forma	236.37	345.62	-0.68389	0.49405
Solido	17.158	345.62	0.93856	0.49405
Extensión	-11.104	5.6179	-1.9765	0.048093
Energía	-63.555	59.164	-1.0742	0.28273
Contraste	-0.11492	0.24984	-0.45996	0.64554
Correlación	-4.298	8.5004	-0.50562	0.61312
Varianza	0.48111	0.21694	2.2177	0.026574
Homogeneidad	56.007	51.195	1.094	0.27396
Suma Porcentaje	2.9519	1.4975	-1.9712	0.0487
Suma de Varianzas	-0.12115	0.055435	-2.1854	0.028858
Suma de Entropías	-17.672	14.393	-1.2279	0.2195
Entropía	5.113	9.6268	0.53113	0.59533
Diferencias de varianzas	0	0	NAN	NAN
Diferencias de entropía	-0.81961	10.835	0.075642	0.9397
Correlación1	37.129	26.392	1.4068	0.15948
Correlación2	14.366	10.148	1.4156	0.15688

5. Conclusiones

Esta investigación se implementan un modelo de clasificación y selección para el análisis de cáncer en una imagen mastográfica.

El estimador de mínimos cuadrados fue implementado para ajustar el modelo y partir del criterio AUC se obtiene la clasificación de tejidos sanos y cancerosos.

Para evaluar la eficiencia de técnicas de selección, se implementaron tres técnicas paso-paso, selección hacia atrás y hacia adelante. Los resultados experimentales demostraron que un modelo con menos característica puede predecir tumores benignos y malignos con excelente rendimiento.

Además, también se evaluó la efectividad de cada uno de los modelos implementados para selección de características, y con base a evidencia estadística el mejor algoritmo es el paso-paso.

Para trabajos futuros, se realizará el diseño de un biomarcador con base a los resultados obtenidos en este trabajo para la selección de características.

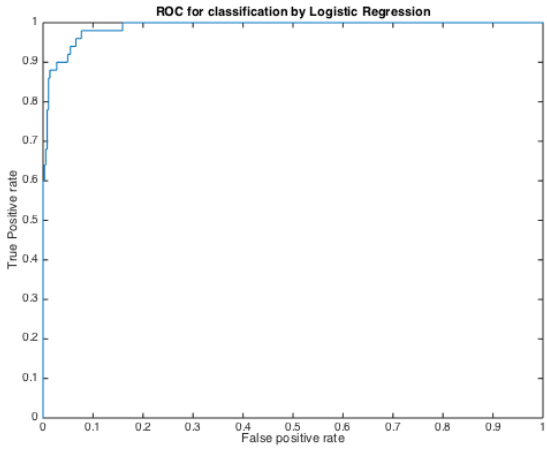

Fig. 4. AUC 0.8554 Paso-paso

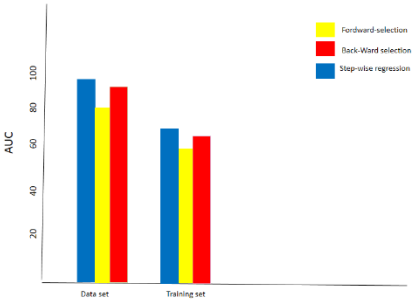

Fig. 5. Comparación de porcentajes AUC para cada modelo de selección

Tabla 4. Datos estadísticos de cada una de las técnicas de selección

	Paso-paso	Hacia-adelante	Hacia atrás
Datos de prueba	13.9	22.5	60.7
Datos de entrenamiento	19.2	19.6	24.6

6. Reconocimientos

El proyecto de investigación pudo realizarse gracias a la colaboración entre la Universidad Autónoma de Zacatecas y al grupo de Investigadores del Doctorado en Ciencias de la ingeniería.

Referencias

[1] J. Ferlay, M. Colombet, I. Soerjomataram, C. Mathers, D. M. Parkin, M. Piñeros, A. Znaor, and F. Bray. Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods, 2019.

[2] World Health Organization (WHO). Cancer burden rises to 18.1 million new cases and 9.6 million cancer deaths in 2018. International Agency for Research on Cancer, 2018.

[3] J. Ferlay, M. Colombet, I. Soerjomataram, C. Mathers, D. M. Parkin, M. Piñeros, A. Znaor, and F. Bray. Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods, 2019.

[4] Mette Kalager, Marvin Zelen, Frøydis Langmark, and Hans Olov Adami. Effect of screening mammography on breast-cancer mortality in Norway. New England Journal of Medicine, 2010.

[5] Heidi D. Nelson, Rochelle Fu, Amy Cantor, Miranda Pappas, Monica Daeges, and Linda Humphrey. Effectiveness of breast cancer screening: Systematic review and meta-analysis to update the 2009 U.S. Preventive services task force recommendation, 2016.

[6] Michael Marmot, D. G. Altman, D. A. Cameron, J. A. Dewar, S. G. Thompson, and Maggie Wilcox. The benefits and harms of breast cancer screening: An independent review, 2012.

[7] T. W. Freer and M. J. Ulissey. Screening mammography with computer aided detection: Prospective study of 12,860 patients in a community breast center. Radiology, 2001.

[8] K. H. Ng and M. Muttarak. Advances in mammography have improved early detection of breast cancer, 2003

[9] Jessica W.T. Leung, Frederick R. Margolin, Katherine E. Dee, Richard P. Jacobs, Susan R. Denny, and John D. Schrumpf. Performance parameters for screening and diagnostic mammography in a community practice: Are there differences between specialists and general radiologists? American Journal of Roentgenology, 2007.

[10] Kevin C. Oeffinger, Elizabeth T. H. Fontham, Ruth Etzioni, Abbe Herzig, James S. Michaelson, Ya-Chen Tina Shih, Louise C. Walter, Timothy R. Church, Christopher R. Flowers, Samuel J. LaMonte, Andrew M. D. Wolf, Carol De Santis, Joannie Lortet-Tieulent, Kimberly Andrews, Deana Manassaram-Baptiste, Debbie Saslow, Robert A. Smith, Otis W. Brawley, and Richard Wender. Breast Cancer Screening for Women at Average Risk. JAMA, 2015.

[11] Syed Jamal Safdar Gardezi, Ahmed Elazab, Baiying Lei, and Tianfu Wang. Breast cancer detection and diagnosis using mammographic data: Systematic review, 2019.

[12] Yanfeng Li, Houjin Chen, Gustavo Kunde Rohde, Chang Yao, and Lin Cheng. Texton analysis for mass classification in mammograms. Pattern Recognition Letters, 2015.

[13] Yuzheng Wu, Maryellen L. Giger, Kunio Doi, Carl J. Vyborny, Robert A. Schmidt, and Charles E. Metz. Artificial neural networks in mammography: Application to decision making in the diagnosis of breast cancer. Radiology, 1993.

[14] Mohamed Meselhy Eltoukhy and Ibrahima Faye. An adaptive threshold method for mass detection in mammographic images. In IEEE ICSIPA 2013 - IEEE International Conference on Signal and Image Processing Applications, 2013.

[15] Jonathan Hernández-Capistrán, Jorge F. Martínez-Carballido, and Roberto Rosas-Romero. False Positive Reduction by an Annular Model as a Set of Few Features for Micro-calcification Detection to Assist Early Diagnosis of Breast Cancer. Journal of Medical Systems, 2018.

[16] M. A. Domínguez, M. Marcos, R. Meirino, E. Villa franca, M. T. Dueñas, F. Arias, and E. Martínez. Prognostic and predictive factors in early breast cancer, 2001.

[17] Utkarsh Mahadeo Khaire and R. Dhanalakshmi. Stability of feature selection algorithm: A review, 2019.

[18] Daniel C Moura and Miguel A Guevara López. An evaluation of image descriptors combined with clinical data for breast cancer diagnosis. International journal of computer assisted radiology and surgery, 8(4):561–574, 2013.