



UNIVERSIDAD DE INVESTIGACIÓN DE TECNOLOGÍA EXPERIMENTAL YACHAY

Escuela de Ciencias Matemáticas y Computacionales

TÍTULO: AUTOMATED IDENTIFICATION OF BREAST CANCER USING DIGITALIZED MAMMOGRAM IMAGES

Trabajo de integración curricular presentado como requisito
para la obtención del título de Ingeniero en Tecnologías de
Información.

Autor:
Chachalo Gómez Bryan Patricio

Tutor:
PhD. Oscar Guillermo Chang Tortolero

Urcuquí – marzo de 2021

Urcuquí, 28 de mayo de 2021

SECRETARÍA GENERAL
(Vicerrectorado Académico/Cancillería)
ESCUELA DE CIENCIAS MATEMÁTICAS Y COMPUTACIONALES
CARRERA DE TECNOLOGÍAS DE LA INFORMACIÓN
ACTA DE DEFENSA No. UITEY-ITE-2021-00012-AD

A los 28 días del mes de mayo de 2021, a las 16:30 horas, de manera virtual mediante videoconferencia, y ante el Tribunal Calificador, integrado por los docentes:

Presidente Tribunal de Defensa	Dr. CUENCA PAUTA, ERICK EDUARDO , Ph.D.
Miembro No Tutor	Dr. INFANTE QUIRPA, SABA RAFAEL , Ph.D.
Tutor	Dr. CHANG TORTOLERO, OSCAR GUILLERMO , Ph.D.

El(la) señor(ita) estudiante **CHACHALO GOMEZ, BRYAN PATRICIO**, con cédula de identidad No. 1004147961, de la **ESCUELA DE CIENCIAS MATEMÁTICAS Y COMPUTACIONALES**, de la Carrera de **TECNOLOGÍAS DE LA INFORMACIÓN**, aprobada por el Consejo de Educación Superior (CES), mediante Resolución **RPC-SO-43-No.496-2014**, realiza a través de videoconferencia, la sustentación de su trabajo de titulación denominado: **AUTOMATED IDENTIFICATION OF BREAST CANCER USING DIGITIZED MAMMOGRAM IMAGES**, previa a la obtención del título de **INGENIERO/A EN TECNOLOGÍAS DE LA INFORMACIÓN**.

El citado trabajo de titulación, fue debidamente aprobado por el(los) docente(s):

Tutor	Dr. CHANG TORTOLERO, OSCAR GUILLERMO , Ph.D.
--------------	--

Y recibió las observaciones de los otros miembros del Tribunal Calificador, las mismas que han sido incorporadas por el(la) estudiante.

Previamente cumplidos los requisitos legales y reglamentarios, el trabajo de titulación fue sustentado por el(la) estudiante y examinado por los miembros del Tribunal Calificador. Escuchada la sustentación del trabajo de titulación a través de videoconferencia, que integró la exposición de el(la) estudiante sobre el contenido de la misma y las preguntas formuladas por los miembros del Tribunal, se califica la sustentación del trabajo de titulación con las siguientes calificaciones:

Tipo	Docente	Calificación
Miembro Tribunal De Defensa	Dr. INFANTE QUIRPA, SABA RAFAEL , Ph.D.	9,0
Tutor	Dr. CHANG TORTOLERO, OSCAR GUILLERMO , Ph.D.	9,0
Presidente Tribunal De Defensa	Dr. CUENCA PAUTA, ERICK EDUARDO , Ph.D.	9,5

Lo que da un promedio de: **9.2 (Nueve punto Dos)**, sobre 10 (diez), equivalente a: **APROBADO**

Para constancia de lo actuado, firman los miembros del Tribunal Calificador, el/la estudiante y el/la secretario ad-hoc.

Certifico que *en cumplimiento del Decreto Ejecutivo 1017 de 16 de marzo de 2020, la defensa de trabajo de titulación (o examen de grado modalidad teórico práctica) se realizó vía virtual, por lo que las firmas de los miembros del Tribunal de Defensa de Grado, constan en forma digital.*

CHACHALO GOMEZ, BRYAN PATRICIO
Estudiante

Dr. CUENCA PAUTA, ERICK EDUARDO , Ph.D.
Presidente Tribunal de Defensa

ERICK EDUARDO CUENCA PAUTA
Digitally Signed by ERICK
CUENCA PAUTA
Date: 2021.06.04 17:52:16
+0500

Dr. CHANG TORTOLERO, OSCAR GUILLERMO , Ph.D.
Tutor



Dr. INFANTE QUIRPA, SABA RAFAEL , Ph.D.
Miembro No Tutor

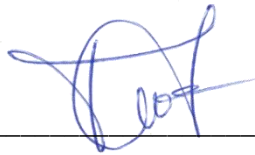
SABA RAFAEL
INFANTE
QUIRPA
Firma digitalmente por
TATIANA
BEATRIZ
TORRES
MONTALVÁN
Fecha: 2023.03.03
15:02:14 -0500

TORRES MONTALVÁN, TATIANA BEATRIZ
Secretario Ad-hoc

AUTORÍA

Yo, **BRYAN PATRICIO CHACHALO GOMEZ**, con cédula de identidad **1004147961**, declaro que las ideas, juicios, valoraciones, interpretaciones, consultas bibliográficas, definiciones y conceptualizaciones expuestas en el presente trabajo; así cómo, los procedimientos y herramientas utilizadas en la investigación, son de absoluta responsabilidad de el/la autora (a) del trabajo de integración curricular, ensayo o artículo científico. Así mismo, me acojo a los reglamentos internos de la Universidad de Investigación de Tecnología Experimental Yachay.

Urcuquí, marzo de 2021.



Bryan Patricio Chachalo Gómez

CI: 1004147961

AUTORIZACIÓN DE PUBLICACIÓN

Yo, **BRYAN PATRICIO CHACHALO GOMEZ**, con cédula de identidad **1004147961**, cedo a la Universidad de Tecnología Experimental Yachay, los derechos de publicación de la presente obra, sin que deba haber un reconocimiento económico por este concepto. Declaro además que el texto del presente trabajo de titulación no podrá ser cedido a ninguna empresa editorial para su publicación u otros fines, sin contar previamente con la autorización escrita de la Universidad.

Asimismo, autorizo a la Universidad que realice la digitalización y publicación de este trabajo de integración curricular, en el repositorio virtual, de conformidad a lo dispuesto en el Art. 144 de la Ley Orgánica de Educación Superior. En caso que el ensayo o artículo se encuentre aprobado para publicación en bases indexadas, únicamente se publicará el resumen del mismo.

Urcuquí, marzo de 2021.



Bryan Patricio Chachalo Gómez

CI: 1004147961

Dedicatoria

Con todo mi ser, a mi madre Mila Gómez.

Bryan Patricio Chachalo Gómez

Agradecimiento

Quiero expresar mi más profundo agradecimiento a Lorena Guachi, notable mentora y excelente persona. “Profe, al final de todo ya habíamos acabado la tesis.”

A todos los profesores que formaron parte de mi formación en Yachay Tech, en especial a Lorena Guachi, Saravana Prakash, Rigoberto Fonseca, y Oscar Chang; gracias por la confianza, apoyo y por haber compartido conmigo sus conocimientos.

A mis amigos los cuales evocan alegría, tristeza, recuerdos de las horas sin dormir y mucho más. Sin dudar la nostalgia está presente al momento de escribir estos párrafos al saber que la vida nos puede llevar lejos de los caminos que tomemos, pero me reconforta saber que los lazos que formamos perduraran entre todos.

Finalmente agradezco a mi toda mi familia, soporte fundamental en lo que fue mi vida universitaria. A mis hermanos, Kimberly, David y Benjamin por ayudarme en aquellos momentos de necesidad. A mis padres Mila Gómez y Segundo Chachalo por todo su apoyo y brindado a lo largo de mi vida y por darme la oportunidad de estudiar esta carrera.

¡Al fin acabe la U mami!

Bryan Patricio Chachalo Gómez

Resumen

*El cáncer afecta a cualquier órgano invadiendo y extendiéndose de forma incontrolada por el cuerpo. Según la Organización Mundial de la Salud, el cáncer de mama es uno de los principales cánceres que afectan a las mujeres de todo el mundo. El tratamiento oportuno de quienes desarrollan cáncer mejora el pronóstico de esta enfermedad e incluso salva vidas. Sin duda, en el diagnóstico del cáncer, la clasificación adecuada de los carcinomas en benignos, malignos y normales es una tarea compleja. Se presenta un algoritmo basado en el Diagnóstico Asistido por Ordenador (CAD) para detectar el cáncer de mama mediante mamografías. En esta implementación del CAD, se utilizan transformaciones como la binarización, el suavizado de umbrales y la operación principal, la onda de Gabor, para el preprocesamiento de las mamografías con el fin de suprimir etiquetas e información innecesaria y obtener las mejores características para la clasificación. Utilizamos técnicas como el análisis de componentes principales (PCA por sus siglas en inglés), *t*-distributed stochastic neighbor embedding (TSNE) y una colección de modelos de varianza estadística para identificar y reducir el espacio de características encontradas. Por último, se hace uso de la técnica de *k*-Nearest Neighbors para la clasificación (*k*-NN) del cáncer.*

Palabras clave: *Detección de cáncer de mama, aprendizaje automático, procesamiento de imágenes*

Abstract

Cancer affects any organ uncontrollably invading and spreading along the body. According to World Health Organization breast cancer is on top of the leading cancers in affecting women around the world. Early treatment of people who develop cancer improves the prognosis of this disease and even saves lives. Unquestionably, in cancer diagnosis, the proper classification of carcinomas into benign, malignant and normal is a complex undertaking. An algorithm based on Computer Aided Diagnosis (CAD) is presented to detect breast cancer using mammograms. In this CAD implementation, transformations such as binarization, threshold smoothing and the main operation, Gabor wavelet, are used for preprocessing to suppress unnecessary labels and information and to obtain the best identifying features. We use techniques such as principal component analysis (PCA), t-distributed stochastic neighbor embedding (TSNE), and a collection of statistical variance models to identify features and reduce the feature space. Finally, we examine the k-Nearest Neighbors technique for classification (k-NN).

Keywords: *Breast Cancer Detection, Machine Learning Classifiers, Image Processing*

Content

Content.....	1
1. Introduction.....	2
2. Objectives	3
3. State-of-the-Art	3
4. Materials and Methods.....	4
4.1. Dataset.....	4
4.2. Method	5
4.3. Pre-processing	6
4.4. Feature Extraction.....	7
4.5. Dimensionality Reduction (DR).	7
4.6. Feature Selection	7
4.7. Classification	8
5. Results and Discussion.....	9
6. Conclusion	12
References.....	13

1. Introduction

Breast cancer is the leading cause of death in women worldwide, and most cases are diagnosed at advanced stages [1]. Early diagnosis, by providing timely care, improves the chances of survival and is therefore a key health strategy. Currently, breast cancer can be screened, detected or diagnosed with numerous tools and technologies. In this regard, mammography is a costly and time-consuming method and is essentially the only imaging modality widely used for breast cancer screening. Therefore, more effective non-invasive tools are needed. Many approaches have been developed for this purpose, including the use of mammograms to assist radiologists in reading mammograms and diagnosing cancer.

Existing approaches mainly operate on a few common steps, such as preprocessing tasks, feature extraction and classification [2]. Preprocessing is performed to improve the visual quality of mammography and the detectability of breast abnormalities. Feature extraction generates a set of discriminative data that is then used as input for the classification step. The ability to predict the class/category of a given data feature from images is referred to as classification [3, 4].

One of the problems faced by radiologists is that the images acquired by the mammographer are often of low quality; they have a slight dissimilarity between normal, benign and malignant cancer tissues leading to inaccurate results. The digitized images taken by the mammographer need to be improved, because the image features can be distinguished and can reflect subtle variation in the order of many degrees. Therefore, preprocessing tasks for mammograph image enhancement become critical before feature extraction. For image enhancement, some works based on spatial and frequency filtering, interpolations and even artificial intelligence techniques have been proposed. For example, histogram equalization (HE) is used as one of the most popular methods for contrast enhancement, which modifies the histogram of gray levels of an image to a uniform distribution [5]. But in many cases, it produces an over enhancement in the output image and a loss of local information which is the information of the values of the pixels that lie in a neighborhood of a given pixel's location.

Models such as LCM-CLAHE [6] are proposed to overcome this limitation. This model performs optimal contrast without losing local information, distance and angle between points, in the mammography image. LCM-CLAHE consists of two processing stages to increase the potential of contrast enhancement and also to preserve local details in the image. In addition, mathematical algorithms as cubic, nearest-neighbor, and linear interpolations are also exploited to reconstructed images degraded by noise or blur effect [7, 8]. Another recognized algorithm for improving image quality is Bottom-hat [9] technique, this filter enhances black spots in a white background. General uncertainty relations limit the resolution of two-dimensional spatial linear filters for orientation, spatial frequency, and 2D spatial position [10]. A family of optimal 2D filters, whose spatial weighting functions are generated by exponentiated bivariate second-order polynomials with complex coefficients, achieves the theoretical lower limit for the joint entropy or uncertainty of these variables.

To address the issue of low detection accuracy in breast cancer due to poor mammography image quality; we proposed a novel method for clearly distinguishing three different breast conditions by using pre-processing techniques prior to feature extraction. The approach presented in this paper is designed and implemented using the Mammographic Image Analysis Society (MIAS) database [11], as better

resolution is obtained than the images that are enhanced in the DDSM approach [12]. In addition, we used the Gabor wavelet to clearly distinguish between normal and abnormal tissues in digital mammograms. The goal of Gabor's wave is to use elliptic generalization of unidimensional elemental functions [13]. It should also be noted that the Gabor filter bank with different orientations and scales disclosed in this work extracts texture patterns such as edges, lines, spots, and flat areas in images, which aids in the differentiation of normal and malignant tissues. Finally, the k-NN classifier for classification is discussed.

2. Objectives

Aim:

To implement a CAD system methodology to assess the contribution of morphological operations and enhancement of features like the Gabor wavelet on mammograms in binary (cancer, non-cancer) and non-binary (normal, benign, malignant) classification.

Objectives:

- Data sets will be collected from the UK-based Mammographic Image Analysis Society, which maintains the digital mammogram MIAS database.
- Noise in mammograms will be suppressed using morphological operations like erosion, dilation, thresholding, and binarization, as well as the Gabor wavelet for enhancement.
- Mammogram features will be extracted from the entire array, with the ones that contribute the most to the predictor variable being kept. For this, techniques like PCA and t-SNE, as well as the ANOVA F-test, will be used to exclude the characteristics that are unrelated to the target variable.
- The results of binary and no binary classification will be obtained using the k-NN method to see whether there is a correlation between the use of preprocessing techniques and mammogram classification.

3. State-of-the-Art

This section gives an overview of newly developed computer-aided diagnosis (CAD) tools for mammogram-based breast cancer detection.

Raghavendra [2] developed a CAD system for mammogram classification using Gabor wavelet technique to enhance mammogram features, Local Sensitive Discriminant Analysis (LSDA) to reduce the dimensional space of the feature matrix, and used machine learning techniques for classification. In Raghavendra's study, he tested classification methods such as decision tree, LDA, k-NN, QDA, SVM, AdaBoost and Fuzzy, and found that k-NN outperforms all of them. The achieved accuracy in Raghavendra paper was 98.69%, sensitivity of 99.34% and specificity of 98.26% in k-NN classifier using 690 mammogram images of the DDSM [12] database. Arfan [14] has developed a framework using a combination of deep Convolutional Neural Network (CNN) with Support Vector Machine (SVM). This method performs preprocessing and enhancement quality of the images, using Deep Convolutional Neural Network (CNN) for features extraction and performs classification with Support Vector Machine (SVM). Arfan's proposed framework has attained an accuracy of 93.35% and 93% sensitivity using the standard dataset MIAS and DDMS. Alkhaleefah [15] proposes a one-class classification (normal versus abnormal). The Alkhaleefah approach combines deep learning and transferring learning. His research

focuses on the principle of transfer learning, in which the power of a Convolutional Neural Network (CNN) can be used as a features extractor to assist in the classification of benign and malignant breast cancer images. This method is divided into three steps. In phase one, he used a CNN trained on spine MRI images. With a few benign and malignant breast pictures, the CNN was fine-tuned and retrained in phase 2. In phase 3, Alkhaleefah fed an RBF-Based SVM with 92.0%, 86.0% precision, and 100% sensitivity using the learned features from phase 2. Tariq et al. [16] developed a CAD system for the detection of breast cancer using mammography images. This CAD system extracts largely discriminating features on the global level for representation and texture characteristics using co-occurrence matrices calculated via the single offset vector. Tariq uses Multilayer perceptron neural network with optimized architecture and fed it with individual feature sets. Using the mini-MIAS database creates his training, cross-validation, and test data. This CAD system achieved an accuracy higher than 99% for both target categories (normal and malignant). Hussain et al. [17] have employed Support Vector Machine (SVM) kernels and Decision Tree to distinguish cancer mammograms from normal subjects. They proposed the use of features such as texture, morphological entropy-based, scale-invariant feature transform (SIFT), and elliptic Fourier descriptors (EFDs). Using Jack-knife 10-fold cross-validation they fed the ML classifiers with the proposed features. Evaluating the performance in terms of specificity, sensitivity, Positive predictive value (PPV), negative predictive value (NPV), false-positive rate (FPR), and receive operating curve (ROC). Obtaining the highest performance based on a single feature extracting strategy using Bayesian approach with texture and EFDs features, and SVM RBF and Gaussian kernels with EFDs features whereas highest AUC with a single feature was obtained using Bayesian approach by extracting texture, morphological, EFDs and entropy features and SVM RBF and Gaussian kernels with EFDs features. According to Hussain's findings, various machine learning techniques perform better with different features extracted using different feature extraction strategies.

Researchers have developed CAD methods for the classification of normal, benign, and malignant mammograms, as seen in the study above. In addition, most of the papers used a small sample size and only considered two cases for classification. This paper proposes a three-class (normal, benign, and malignant) and two-class (cancer, non-cancer) CAD classification system in order to create a system for breast cancer diagnosis with a data collection. This system proposes the use of morphological operations for focusing on the region of interest ROI (the breast), feature extraction and enhancement techniques, feature matrix reduction techniques, and selecting the features that contribute the most to mammogram classification.

4. Materials and Methods

4.1. Dataset

Machine learning is a popular artificial intelligence technique in intelligent systems such as computer vision, language processing and classification. Numerous mammography image databases exist and the selection of the database to be used in a machine learning process is a critical step in system design and implementation. The Mammographic Image Analysis Society is a UK-based research organization that maintains the MIAS database of digital mammograms [18], which has a resolution of 1024 pixels. It is a collection of 322 digitized films that includes normal, benign, and malignant cases and radiologist's "ground truth" with the detection of any abnormalities that may be present. Total of 330 samples were obtained from MIAS dataset for training (207 normal, 69 malignant, 54 benign), while 492 were used for testing (158 normal, 231 malignant, 103 benign) purposes. In order to

provide an effective breast cancer diagnosis classification, it is essential the use of mammogram dataset that contains close number of images for each lesion/class. By flipping images horizontally and vertically, we were able to increase the number of images for each case (207 normal, 207 malignant, 207 benign) using data augmentation. Data augmentation process was performed on the mammograms randomly until the number of 207 mammograms was reached for each case. Following this, the database was randomly divided for the subsequent processes of preprocessing and/or classification.

4.2. Method

The approach in this paper has the main objective of classifying mammogram images in three condition types: normal, benign, and malignant. The main work flow is shown in Fig. 1. The input mammography images are enhanced in the preprocessing phase to improve their quality and remove unwanted information. Feature extraction then extracts meaningful data to distinguish three different conditions from the mammography image. After, dimensionality reduction techniques are applied, to reduce the number of discriminative features. Besides, Analysis of Variance (ANOVA f-test) is used to select the most remarkable features from the previous stage. Finally, classification stage is done by using supervised classifiers independently, in order to determine their performance on mammogram image classification.

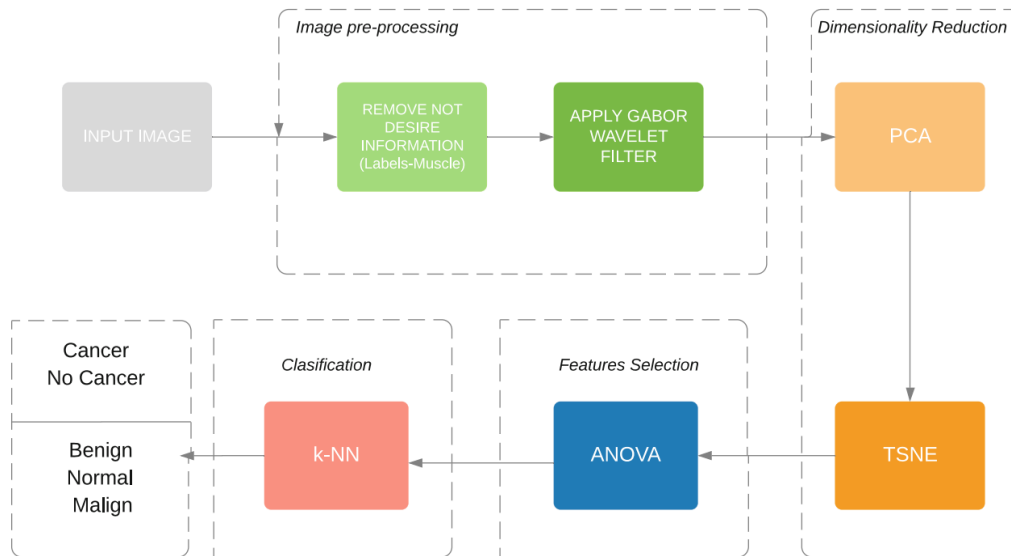


Fig 1. Description of the methodology.

Algorithm 1 Remove Labels

```

1: function REMOVELBL(Img, ImgBin, color)
2:   height = height.Img
3:   width = width.Img
4:   for i = 0 to height do
5:     for j = 0 to width do
6:       if color = 0 then
7:         if ImgB[i, j] = 0 then
8:           Img[i, j] = 0
9:         else if ImgB[i, j] = 255 then
10:          Img[i, j] = 0
11:        end if
12:      end if
13:    end for
14:  end for
15: end function

```

a)

Algorithm 2 Binarize Through a ROI

```

1: function THRESHROI(Img, Thr)
2:   height = height.Img
3:   width = width.Img
4:   ImgRes = ZeroMatrix(height, width)
5:   Left = width.Img/2
6:   Rigth = width.Img/2
7:   for i = 0 to height do
8:     for j = 0 to width do
9:       if j < Left or j > Rigth then
10:        if Img[i, j] > Thr then
11:          ImgRes[i, j] = 255
12:        else
13:          ImgRes[i, j] = 0
14:        end if
15:      end if
16:      ymidI = 1
17:      ymidR = 1
18:    end for
19:  end for
20: end function

```

b)

Algorithm 3 Remove Edges

```

1: function RMVEDGES(Img)
2:   Ret = Threshold(Img, 240, 255, BINARY)
3:   edgesTh = Threshold(Img, 240, 255, BINARY)
4:   Kernel = OnesMatrix((20, 20), uint8)
5:   dilEdges = Dilate(edgesTh, kernel, iter = 1)
6:   NoEdges = RemoveLbl(Img, dilEdges, 255)
7: end function

```

c)

Algorithm 4 Generate Filter Gabor Wavelet

```

1: function BUILDFILTER(Img)
2:   height = height.Img
3:   width = width.Img
4:   ImgRes = ZeroMatrix(height, width)
5:   Filter(Img)
6:   for Kernel = 0 to Filter do
7:     fimg = Filter2D(Img, CV_8UC3, Kernel)
8:     Maximum(accum, fimg, accum)
9:   end for
10:
11: end function

```

d)

Algorithm 5 Pre Processing Images

```

1: function PREPROCESSING(Img)
2:   Img = Resize(Img, (1360, 796))
3:   Smooth = GaussianBlur(Img, (5, 5), 0)
4:   Ret = Threshold(Img, 65, 255, BINARY)
5:   SmoothBin = Threshold(Img, 65, 255, BINARY)
6:   kernel = OnesMatrix((55, 55), uint8)
7:   erosion = Erode(SmoothBin, kernel, iter = 1)
8:   dilation = Dilate(SmoothBin, kernel, iter = 1)
9:   NoLbl = RemoveLbl(Img, dilation, 0)
10:  smoothNoLbl = GaussianBlur(NoLbl, (39, 39), 0)
11:  smoothRoi = ThreshROI(smoothNoLbl, 150)
12:  NoLbl2 = RemoveLbl(NoLbl, smoothRoi, 255)
13: end function

```

e)

Fig 2 The pseudo-code of the proposed method. a) Remove labels; b) ROI binarization; c) Remove edges; d) Build Gabor filter; e) Pre-processing function.

4.3. Pre-processing

At this stage, unwanted data is removed: labels, margins and pectoral tissues, which may degrade the accuracy of the proposed approach. When using data augmentation, mammograms with different resolutions are produced since some techniques involve cutting the images; thus, the sequence of preprocessing techniques begins with resizing to 1024 x 1024 pixels for uniformity. Then, a Gaussian filter with a kernel of 5x5 is applied. Following that, images are binarized with global thresholding Th1=65. It is followed by Erosion and then dilation with a kernel size of 55x55. Smoothing is then applied to the images using a kernel of size 39x39, and a second binarization is applied across a region of interest using a threshold of Th2 = 150.

Since each mammogram is unique, the parameterization of the morphological processes, as well as the order in which they are used, tend to differ. As a result, a trial-and-error search of the order of the morphological processes applied, as well as the parameterization of each one, was conducted, with the best parameterization values for the entire set of mammograms being those previously defined in this section. These parameter values are set for the entire data set. Figure 2a, b, c explains in detail the order of operations performed during this preprocessing process in pseudocode format.

4.4. Feature Extraction

It consists in the enhancement and extraction of features from the images. The Gabor wavelet has been widely used in image processing research, and its tunable parameters are critical to its effectiveness in applications such as facial expression classification, Gabor networks for face reconstruction, fingerprint recognition, and others [19]. Therefore, various Gabor wavelet function value combinations have been tested, with the best ones listed in Table 1. In this work, Gabor wavelet filters were applied over the entire image to extract discriminative features. Then, all pixels in the image were fused into a 32-bit floating point.

4.5. Dimensionality Reduction (DR).

Since feature extraction usually generates massive amounts of data that are difficult to analyze, this stage seeks a low-dimensional representation of the feature matrix obtained in the previous stage. The size of a feature matrix is $N \times M$, where N is the total number of feature sets in the dataset and M is the number of mammography image samples. In this approach, Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (TSNE) are used to reduce the dimensionality of mammography. Following the description of t-SNE provided by the authors in [20], the t-SNE algorithm converts data into a lower-dimensional space, where a small gap between two points indicates that they were nearby in the original space. In comparison, if two points were apart in the original data set, they would be far apart after applying PCA.

The theoretical, computational, and empirical differences between the two methods indicate that one is not necessarily "better" than the other. However, the PCA's complexity is $\mathcal{O}(p^2n + p^3)$, and considering p fixed, this equals $\mathcal{O}(n)$. On the other hand, t-SNE computational and memory complexity are both $\mathcal{O}(n^2)$ [20]. Since t-SNE scales quadratically in the number of objects N , its applicability is limited to data sets with a few thousands of entry objects, learning becomes too sluggish to be useful after that (and memory requirements become too large). As a result, in this approach, we first used a low-cost method to reduce the dataset, such as PCA, so that t-SNE can be performed on the reduced input data, making it feasible.

PCA has no parameters, while the t-SNE has many; however, default parameters have been used in both cases because the t-SNE output is reasonably resilient to changes. The matrix dimensions at this point are $N \times M$, where N is the total number of mammograms (322), and M is their dimension (1024×1024 corresponding to 1048576 components). To begin, PCA is used to save 95% of the variance, resulting in a matrix dimension of $N = 322$ and M ranging from 8 to 20 components. Then, t-SNE is applied to this new reduced matrix, in the end keeping 2 components. Fig. 3a. shows the implementation.

4.6. Feature Selection

This stage uses analysis of variance (ANOVA f-test) to automatically select, from the results obtained in the previous stage, the relevant characteristics with the greatest contribution to the predictor variable. It helps in enhancing classifier performance, computational time and cost-effective. An F-statistic, also known as an F-test, is a class of statistical tests that use a statistical test like ANOVA to measure the ratio between variance values, such as the variance from two separate samples or the explained and unknown variance. An ANOVA f-test is a form of F-statistic that uses

the ANOVA process. The results of this test are used to pick features from the dataset, with features that are independent of the goal variable being dropped.

The scikit-learn machine library was used in this implementation, which includes an implementation of the ANOVA f-test in the *f_classif()* function [21]. The *SelectKBest* class was used in a feature selection strategy to select the top k most important features (largest values). We had to use repeated stratified k-fold cross-validation to test model configurations on classification tasks in order to pick a good number of features. The *RepeatedStratifiedKFold* class was used to perform three 10-fold cross-validation repeats. Following this grid scan, we discovered that the best number of selected features in this case is 17. Fig. 3b. shows the flow of data through the selection of characteristics using scikit-learn ANOVA f-test.

Algorithm 6 Dimensionality Reduction

```

1: function DIMREDUC(Img)
2:   pca = PCA(n_components = 0.95)
3:   ImgReduced = PCA.FitTransform(Img)
4:   tsne = TSNE(n_jobs = 150, n_components =
      3, random_state = 42)
5:   TSNEImg = TSNE.FitTransform(ImgReduced)
6: end function

```

a)

Algorithm 7 Image Classification

```

1: function IMGCLA(TSNEImg)
2:   Fvalue = SelectKBest(Fclassif, k = 3)
3:   Xkbest = Fvalue.FitTransform(TSNEImg)
4:   knnReg = neighbors.KNeighborsClassifier()
5:   knnSearch = RandomizedSearchCV(knnReg)
6:   knnSearch = knnSearch.Fit(XTrainImgs, YTrainImgs)
7:   Prediction = CrossValPredict(knnSearch, XTestImgs, YTestImgs)
8: end function

```

b)

Algorithm 8 Main Function

```

1: function MAIN
2:   Img = ReadImage
3:   Img = Preprocessing(Img)
4:   filters = BuildFilters()
5:   resImg = Process(Img, filters)
6:   PPI = RmvEdges(resImg)
7:   dimReducImg = DimReduc(PPI)
8:   clasfImg = ImgCla(dimReducImg)
9: end function

```

c)

Fig 3. The pseudo-code of the proposed method. a) Dimensionality reduction; b) Image classification; c) Main computational calls.

4.7. Classification

This stage explains how to apply a mammography condition to the input pattern by using k-nearest neighbors (k-NN), one in every of the foremost common machine learning strategies. It is based on instances and permits the classification of the latest parts by calculating their distance to any or all the opposite parts $dist(X1, X2)$. The proper functioning of the algorithm depends on the choice of the distance function used and the value of the parameter k , that represents the number of near neighbors to the query x_q . The neighbors are weighted by the distance separating them from the new elements being classified. The effect of noise in classification is reduced when the value chosen for k is greater, but this makes less distinguishable in limits that fall among the classes. K-NN is effective on noisy training data and suitable for cases of a large number of training samples, however, the computation time increases as we need to compute the distance from each instance to all training samples. This work uses the Minkowski distance to achieve higher accuracy with minimal effect due to

the variation of the k parameter [22]. The parameter settings of K-NN are specified in Table 1. The main code sequences of K-NN are shown in Fig. 4 and in Fig. 3c. shows the data flow from mammogram reading to the classification phase using the features after reduction, selection operations.

Table 1. Parameter values.

	Parameter	Tested values	Best value
Gabor kernel	ksize	1; 2; 5; 10; 20; 30; 33; 35; 40; 45	10
	sigma	1; 2; 3; 4; 5; 6; 8; 10; 20	4
	lambda	1; 2; 5; 10; 15; 20	10
	gamma	0.1; 0.2; 0.3; 0.4; 0.5; 0.6; 0.7; 1.0; 1.5	0.5
	psi	0; 0.5; 1; 1.5	0
	ktype	CV_32F	CV_32F
K-NN algorithm	k (number of neighbors)	3; 5; 8; 11; 15	8
	weights	uniform; distance	uniform
	algorithm	auto; ball_tree; kd_tree; brute	auto
	metric	euclidean; manhattan; minkowski	minkowski

Algorithm k-NN algorithm.9

Choose the best values for the parameters using RandomizedSearchCV, prediction with the best_estimator, and/or score.

Input: X: matrix with features to train, y:vector of labels, hyper parameters (n_neighbors, weights, metrics)

- 1: **function** RSEARCHCV(*n_neighbors, weights, metrics*)
 - 2: Create a dictionary with the n_neighbors, weights, metrics values to try.
 - 3: Use Randomized search on hyper parameters
 - 4: **return** best_estimator ▷ best hyper parameters for knn classifier
 - 5: **procedure** KNN_CLASSIFIER(*X_train, y_train, x_test, y_test*)
 - 6: Use RSearchCV to get the best_estimator
 - 7: Fit the model with X_train,y_train
 - 8: Predict labels for x_test
 - 9: **return** predicted values ▷ Score can also be obtained by using y_test
-

Fig 4. The pseudo-code of the k-NN algorithm.

5. Results and Discussion

We used the computational resources of the Quinde 1 supercomputer at the Yachay project in Urcuquí, Ecuador, to process our data. There were 150 high-performance computing cores to increase computational speed (Quinde 1 has 1640 cores and runs Linux Red Hat Enterprise Linux Server version 7.2 (Maipo) little endian). For experimental tests, software routines were implemented in python 3. They use libraries such as *scikit-learn* [21] and execute the portion of machine learning algorithms, and OpenCV [23] for the image processing techniques applied in the pre-processing stage. Similarly, feature enhancements and other Python language tools such as pandas, matplotlib and numpy were used for data handling tasks.

Recall, precision, specificity and overall accuracy defined by equations 1, 2, 3 and 4, respectively, have been evaluated to measure the effectiveness of the proposed method.

Recall, also known as true positive rate (TP), is the proportion of positive cases that were correctly identified:

$$Recall = \frac{d}{c + d} \quad (1)$$

Precision is defined as the proportion of predicted positive cases that were identified as correct, and it can be expressed as follows:

$$Precision = \frac{d}{c + d} \quad (2)$$

Specificity, also known as true negative rate, is the proportion of correctly identified negatives:

$$Specificity = \frac{a}{a + b} \quad (3)$$

Accuracy is defined as the proportion of true and false positives that are correctly identified:

$$Accuracy = \frac{c + d}{a + b + c + d} \quad (4)$$

Where **a** represents the number of correct predictions indicating a negative instance, **b** represents the number of incorrect predictions indicating a positive instance, **c** represents the number of incorrect predictions confirming the negative instance, and **d** represents the number of correct predictions validating the positive instance.

Fig. 5 shows the pre-processing stages applied in the proposed model. It is important to note that the image resulting from the preprocessing was labeled in all its states and show the order of application: (a) Original image, (b) Binarized Image, (c) Erode image, (d) Dilate image, (e) Remove Labels, (f) Smooth Image, (g) Detect and remove muscle, (h) Image applied Gabor wavelet filter, (i) Removing not desire edges and (j) Image applied original pre-processing method, respectively. Looking at the image of the pre-processed mammogram after applying the Gabor filter, i.e., image Fig. 5I, from the above-mentioned sequence of images, differs from the final result of the reference paper [2], as shown in Fig. 5J of [2]. This is only because a different parameter combination was used for the Gabor filter. Particularly, the parameters that were implemented in our proposed model allow us to achieve a greater degree of accuracy in the CAD system by realizing morphological features.

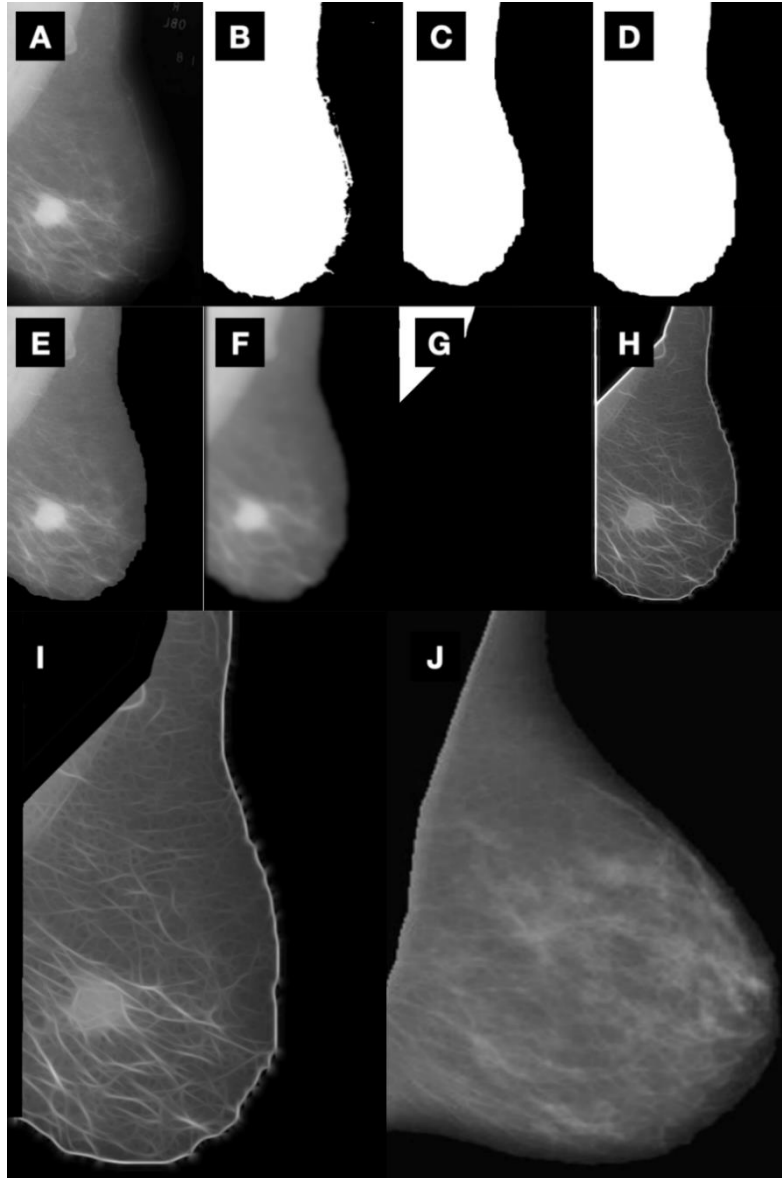


Fig 5. Steps for Pre-processing images: a) Original image; b) Binarized Image; c) Erode image; d) Dilate image; e) Remove Labels; f) Smooth Image; (g) Detect and remove muscle; h) Image applied Gabor wavelet filter; i) Removing not desire edges; j) Image applied original pre-processing method [2].

From our results represented in Table 2, we estimate a value of 98.22% of precision in the classification of the malignant condition, 35.89% in benign condition and 82.14% in normal condition. The precision obtained in the classification of normal cases is because the classifier erroneously mispredicts 8 normal cases as benign cancer cases and 2 as malignant cancer cases. The classifier misclassifies 24 benign cancer cases as normal and 1 case as malignant cancer, resulting in a low percentage of precision in benign cancer prediction. The classifier predicts correctly and does not confound with any of the 3 conditions when it comes to malignant cancer

classification. This shows that the classifier addresses the critical need to predict a malignant cancer case.

While the overall performance values in Table 3 suggest that the selection of a correct classifier is heavily reliant on an accurate evaluation of its performance.

Table 2. Precision and recall values achieved.

		Predicted			Precision
		Normal	Benign	Malignant	
Actual	Normal	46	8	2	82.14%
	Benign	24	14	1	35.89%
	Malignant	0	0	168	98.22%
	Recall	65.71%	64.63%	97.07%	

The examination of all percent accuracy, precision and specificity is very informative and is presented in Table 3. The mean accuracy, a harsh metric for the ten forecasts used in cross-validation, is included in the reported averages. It is worth noting that in Conditions 3 (classification into: normal which does not present any protuberance, benign cancer, and malignant cancer) and 2 (classification into: malignant cancer, benign cancer), averaging is set to total effectiveness, true positives, and true negatives, respectively. The proposed approach has been compared with respect [2] to measure the effectiveness of the introduced pre-processing tasks by using the same classifier. Furthermore, the proposed method has been shown to be more robust than those using deep learning techniques [14, 15]. However, Multilayer Artificial Neural Network (ANN) Mean Square Error [16] achieved the highest overall accuracy of 99.2 percent, but only classifies two conditions (classification into: malignant cancer, benign cancer). Our method can classify in two or three conditions and outperforms in identifying the true positive values.

Table 3. Overall Results

Method	3 Conditions			2 Conditions		
	Accuracy	Precision	Specificity	Accuracy	Precision	Specificity
K-NN (Proposed)	86.99%	77.00%	77.00%	98.48%	98.48%	97.00%
K-NN + Pre-processing (Proposed)	89.39%	80.00%	80.00%	99.18%	99.00%	98.00%
K-NN + Pre-processing [2]	--	--	--	98.69%	99.13%	98.26%
Deep Convolutional Neural Network [14]	--	--	--	93.35%	--	--
Multilayer ANN + Mean Square error [16]	--	--	--	99.20%	--	99.15%
Decision tree classifiers with EFDs [17]	--	--	--	97.22%	--	--
Gaussian with texture feature [17]	--	--	--	97.44%	--	--
AlexNet-Polynomial-Based-SVM [15]	--	--	--	85.00%	--	86.00%
GoogLeNet-CNN-Softmax [15]	--	--	--	77.00%	--	71.00%
Hybrid CNN and RBD-Based SVM [15]	--	--	--	92.00%	--	86.00%

6. Conclusion

The proposed method seeks to automate the classification and segmentation processes in mammography analysis. Normal, benign, and malignant conditions are among the data types to be classified. It is important to note that the preprocessed image, image I in Fig. 5, differs from the final result of the reference work [2], image J in Fig. 5. This is due to the fact that a different parameter combination was used for

the Gabor filter. The parameters used in the proposed model allow a better realization of the morphological characteristics, thus aiming to increase the accuracy of the computer-aided diagnosis system.

Table 3 shows that the proposed method is better in percentage accuracy than the reference method [2], however, it is also observed that the method of [16] outperforms both the proposed and reference method in accuracy and specificity. By examining Table 3 and comparing the findings of [16], it is clear that the features used to feed a classification method, as well as the classification method itself, have a significant impact on the accuracy of the results. Nonetheless, regardless of the classification method used, the use of preprocessing techniques significantly and directly improves the results in the classification phase.

Hence, our model can be improved with automatic parameterization techniques. Focusing on the normal and benign cases, we discovered that both have lower accuracy, which could be improved by: 1. Adding more data to these cases, currently the MIAS database contains 64 benign and 115 normal mammography cases; therefore, having a uniform database would be the first improvement 2. Looking to create new features that will improve the classifier, in classification we used only mammograms; however, other characteristics such as tissue type, abnormality if lumpiness is present, position of lumps, and radius of lumps may be used 3. Modifying the preprocessing stage's parameterization, so far, we have used trial-and-error search for the parameterization in the preprocessing stage; in case, it would be beneficial to introduce a deep learning or artificial intelligence system capable of parameterizing each mammogram case by case, since in this method a parameterization was sought and the results were generalized for the entire mammography database.

The effects of machine learning classifiers in this context could be investigated in the future to determine their overall accuracy in correctly classifying mammography image conditions. Based on the techniques presented in this article, we would also like to have a fully automatic system for image processing and classification of mammography cases. It could assist radiologists in the mammographic interpretation process as an appropriate noninvasive tool.

References

- [1] "World Health Organization Homepage," [Online]. Available: <https://www.who.int/cancer/detection/breastcancer/en/index1.html>. [Accessed 28 May 2019].
- [2] U. Raghavendra, "Application of Gabor wavelet and locality sensitive discriminant analysis for automated identification of breast cancer using digitalized mammogram images," *Appl. Soft Comput.*, vol. 46, pp. 151-161, 2016.
- [3] L. Guachi, G. Robinson, F. Bini and F. Marinozzi, "Automatic colorectal segmentation with convolutional neural network," *Computer-Aided Design and Applications*, vol. 16, no. 5, pp. 836-845, 2019.
- [4] G. Lorena, G. Robinson, P. Stefania, C. Pasquale, B. Fabiano and M. Franco, "Automatic Microstructural Classification with Convolutional Neural Network," in *Information and Communication Technologies of Ecuador (TIC.EC)*, vol. 884, M. Botto-Tobar, L. Barba-Maggi, J. González-Huerta, P. Villacrés-Cevallos, O. S. Gómez and M. I. Uvidia-Fassler, Eds., Cham, Springer International Publishing, 2019, pp. 170-181.

- [5] Y. Yi, B. Zhang, J. Kong and J. Wang, "An improved locality sensitive discriminant analysis approach for feature extraction," *Multimedia Tools and Applications*, vol. 74, no. 1, pp. 85-104, 2015.
- [6] V. Muneeswaran and M. Pallikonda Rajasekaran, "Local contrast regularized contrast limited adaptive histogram equalization using tree and seed algorithm-am aid for mammogram images enhacement," *Satapathy, S.C., Bhateja, V., Das, S.(eds.) Smart Intelligent Computing and Applications. SIST*, vol. 104, pp. 693-701, 2019.
- [7] R. Keys, "Cubic convolution interpolation for digital image processing," *IEEE Trans Acoust Speech Signal Process*, vol. 29, n° 6, pp. 1153-1160, 1981.
- [8] J. Parker, R. Kenyon and D. Troxel, "Comparison of interpolating methods for image resampling," *Medical Imaging, IEEE Transactions on*, vol. 2, no. 1, pp. 31-39, 1983.
- [9] S. Das, "Medical image enhancement techniques by bottom hat and median filtering," *International Journal of Electronics Communication and Computer Engineering*, vol. 5, no. 4, pp. 347-351, 2014.
- [10] H. Meng, "Iris recognition algorithms based on Gabor wavelet transforms," *2006 International Conference on Mechatronics and Automation*, pp. 1785-1789, 2006.
- [11] J. Suckling, "The mammographic image analysis society digital mammogram database," *Digital Mammo*, pp. 375-386, 1994.
- [12] DDSM, "Digital Database for Screening Mammography," [Online]. Available: <https://www.eng.usf.edu/cvprg/Mammography/Database.html>. [Accessed 14 Jun 2019].
- [13] J. Daugman, "Uncertainly relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters," *Journal of the Optical Society of America. A, Optics and image science*, vol. 2, no. 7, pp. 1160-1169, 1985.
- [14] M. Arfan, "Deep Learning based Computer Aided Diagnosis System for Breast Mammograms," *International Journal of Advanced Computer Science and Applications*, vol. 7, pp. 286-290, 2017.
- [15] M. Alkhaleefah and C.-C. Wu, "A hybrid CNN and RBF-based SVM approach for breast cancer classification in mammograms," in *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2018.
- [16] N. Tariq, B. Abid, K. Qadeer, I. Hashim, Z. Ali and I. Khosa, "Breast Cancer Classification using Global Discriminate Features in Mammographic Images," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 2, pp. 381-387, 2019.
- [17] L. Hussain and W. Aziz, "Automated breast cancer detection using machine learning techniques by extracting different feature extracting strategies," *2018 17th IEEE International Conference on Trust, Security and Privacy in Computing and Communications*, pp. 327-331, 2018.
- [18] "The mini-MIAS database of mammograms," [Online]. Available: <https://peipa.essex.ac.uk/info/mias.html>. [Accessed 14 Jun 2019].
- [19] W.-L. Chao, "Gabor wavelet transform and its application," 2010.
- [20] L. van der Maaten and G. Hinton, "Visualizing Data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579-2605, 2008.
- [21] Pedregosa, "Scikit-learn: Machine Learning in Python," *JMLR*, vol. 12, pp. 2825-2830, 2011.

- [22] A. Makandar and B. Halalli, "Pre-processing of Mammography Image for Early Detection of Breast Cancer," *International Journal of Computer Applications*, vol. 144, no. 3, pp. 0975-8887, 2016.
- [23] G. Bradski, "The OpenCV Library," *Journal of Software Tools*, 2000.