A Genetic Programming Approach to Automatically Construct Informative Attributes for Mammographic Density Classification

***Abstract*—Breast density is widely used as an initial indicator of developing breast cancer. At present, current classification methods for mammographic density usually require manual operations or expert knowledge that makes them expensive in real-time situations. Such methods achieve only moderate classification accuracy due to the limited model capacity and computational resources. In addition, most existing studies focus on improving classification accuracy using only raw images or the entire set of original attributes and remain unable to identify hidden patterns or causal information necessary to discriminate breast density classes. It is challenging to find high-quality knowledge when some attributes defining the data space are redundant or irrelevant. In this study, we present a novel attribute construction method using genetic programming (GP) for the task of breast density classification. To extract informative features from the raw mammographic images, wavelet decomposition, local binary patterns, and histogram of oriented gradients have been utilized to include texture, local and global image properties. The study evaluates the goodness of the proposed method on two benchmark real-world mammographic image datasets and compares the results of the proposed GP method with eight conventional classification methods. The experimental results reveal that the proposed method significantly outperforms most of the commonly used classification methods in binary and multi-class classification tasks. Furthermore, the study shows the potential of GP for mammographic breast density classification by interpreting evolved attributes that highlight important breast density characteristics.**

***Index Terms*—genetic programming, attribute construction, breast density, image classification**

I. Introduction

Breast cancer is the most commonly diagnosed cancer with an estimated 2.3 million new cases and $684,996$ deaths worldwide, as per global cancer statistics 2020 [1]. Fig. 1 shows the incidence and mortality rates of the top 10 most common cancers where breast cancer remains prominent. Breast density is an important indicator for developing breast cancer [2]. In medical studies, a woman gets 5-fold increased risk of developing breast cancer if mammographic breast density (MBD) exceeds 75% [3].

Mammography is the most effective technique for breast cancer screening [4]. Radiologists qualitatively evaluate the breast density based on the Breast Imaging and Reporting Data System (BI-RADS) standards [5]. The standard BI-RADS criterion has four categories; BI-RADS I: fatty (0–25%), BI-RADSII: scattered density (26–50%), BI-RADSIII: heterogeneously dense (51–75%), and BI-RADSIV: extremely dense (76–100%) [5]. When radiologists observe mammograms with the naked eye to classify them into BI-RADS categories, there is an obvious inter- and intra-reader variability [2]. In addition, categorizing a large number of mammograms is a tiresome process for the radiologists. Hence, designing and developing a fully automatic classification method to accurately classify mammograms is urgently needed. This requires identification of strategies that capture informative characteristics such as texture, scale, local, global and frequency-based properties, which are conceptually useful to mimic these clinical applications.

When applied to sub-images, feature extraction extracts local features, while when applied to the entire image, it extracts global features. Additionally, features can be extracted from multiple color channels and scales of an image to incorporate texture information. Such classification methods or diagnostic systems are in demand, which can accurately classify a particular breast density level in real-life scenario. In addition to correctly classifying, it is also vital to identify significant features simultaneously that can provide a radiologist with critical visual patterns [6]. In medical domain, identifying a diseased image correctly is more important compared to identifying a non-diseased image correctly [7]. The raw pixels or features extracted directly from mammographic images may not contain enough information to distinguish between images of different breast density levels to ensure accurate classification. Consequently, a number of artifacts can result in redundant or irrelevant extracted features that can affect classification accuracy. This type of problem can be addressed with the use of *feature selection* and *feature construction*, which are mainly used to select important features from the original set, and to construct new informative features based on the original set of features, respectively [8].

During the last decade, deep learning has been actively applied and has provided viable outcomes for breast density classification. Mohamed et al. [9] utlizied AlexNet and transfer learning approach for binary classification of mammograms. They resized the original mammographic images to a reduced size (almost 10 times reduction in size) to suit the architecture of AlexNet. Li et al. [7] utilized ResNet-50 to classify breast density from mammographic images. They could not achieve satisfactory results on a small dataset, as deep learning models usually require large training data to effectively train the classification model.

Genetic Programming (GP) is a bio-inspired evolutionary computation method that automatically evolves models

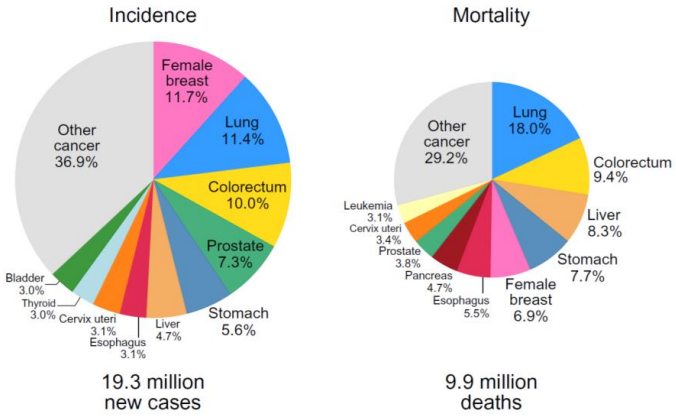


Fig. 1. Distributions of the top 10 most common cancers in 2020 [1].

(computer programs) to solve a problem by using genetic operators such as crossover, mutation, and reproduction [10]. GP explores the search space and has a built-in ability to select prominent attributes with discriminating ability between classes [11]. In addition, the evolved models by GP are used to construct new attributes that can help achieve performance gains [12]. GP can evolve multiple trees in one individual which is termed as multi-tree GP [13]. GP has not only been exclusively utilized for classification, but has likewise been explored extensively for feature selection and construction [11], [14]. In image analysis, GP has been explored in a wide range of applications such as object detection [15], feature extraction [11], feature construction [12], [14], evolving image descriptors [16], and classification [11]. In medical image analysis, GP has been recently explored for skin cancer detection [17], but it has only been studied by Burling et al. [18] for MBD binary classification in fatty and dense categories. They extracted statistical and local binary patterns (LBP) features, and utilized transfer learning in GP for binary image classification. The results show the potential of GP for MBD classification but still need improvement to reach satisfactory results.

Goals: This study aims to develop a novel attribute construction method based on multi-tree GP for breast density classification from mammograms. Different from most existing methods, the proposed method aims at evolving a GP individual with multiple attributes rich in multi-scale, local, and global textural properties provided by frequency-based pyramid-structured wavelet decomposition, LBP, and histogram of oriented gradients (HOG). To better explore the search space, the multiple trees in a GP individual utilize all-index crossover during the evolutionary process where the sub-trees can crossover regardless of their index in a GP individual. By doing so, the proposed method is expected to automatically evolve information-rich attributes with discriminating ability between the MBD image classes. This work aims to address the following research question:

- Can the proposed multi-tree GP approach automatically construct multiple informative attributes to provide better discriminating ability between different breast density levels in mammographic image datasets?
- How well the proposed GP method perform in comparison to the commonly used machine learning classification algorithms: Naïve Bayes (NB), Support Vector Machines (SVM), k-nearest neighbor (k-NN), and decision trees (J48), and ensemble methods: Random Forest (RF), Bagging, Adaboost, and LogitBoost?
- How well does this new method work as compared to existing GP approach(es) for classifying breast density mammograms? and
- Which original attributes are most prominent in providing good classification performance and why?

II. Literature Review

A. Feature Extraction

Feature extraction is used to extract the image features, similar to those visually detected by radiologists, that can accurately characterize mammograms into BI-RADS categories. This study utilizes commonly used statistical features to encompass global image properties and pyramid-structured wavelet decomposition to include multi-scale, local, global, and texture image properties. Moreover, LBP and HOG image descriptors are utilized to include texture and shape image properties. This study uses the following three feature extraction methods:

1) Pyramid Structured Wavelet Decomposition: Texture analysis helps identify the visual characteristics of mammogram tissues which constitutes the basis of clinical diagnosis (BI-RADS) [7]. The pyramid-structured wavelet analysis [19] captures both the local (detailed structure and internal texture) and global (overall properties) information of the mammographic image. We have applied three-level pyramid-structured wavelet decomposition on the original mamographic images as shown in Fig. 2.

To transform the image data to feature vectors, eight statistical features are extracted from each of the original mammogram images. In addition, the eight statistical measures are used to extract informative features from the wavelet coefficients. The statistical features include energy, kurtosis, mean, average energy, standard deviation, entropy, skewness, and norm. Table I shows the mathematical description of these features. More details can be found in [20].

2) Local Binary Patterns: Developed in 1994 [21], LBP is a visual image descriptor widely adopted for texture classification. It computes the labels of the pixels in an image according to the neighborhood of each pixel. By setting each central pixel in a window as a threshold, it compares the pixels surrounding the central pixel and generates the result as a binary number. More details can be found in [21]. In mammographic images, texture of the breast tissue is an important characteristic to classify between different density levels [5].

3) Histogram of Oriented Gradients: It is an image descriptor developed by Dalal et al. [22] which focuses on shape and texture information in image classification. The HOG image descriptor uses magnitude and orientation of the gradient to

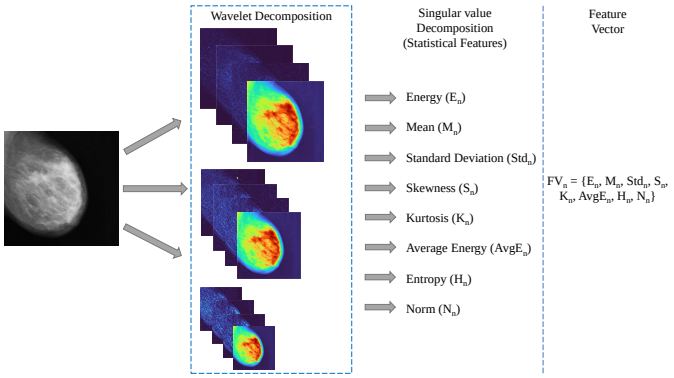


Fig. 2. Converting mammographic images to feature vectors using pyramid-structured wavelet decomposition and statistical features.

TABLE I
STATISTICAL FEATURES

Measure	Mathematical Expression
Energy	$E_{n_i} = \frac{\sum_{j=1}^{J}\sum_{k=1}^{K} x_{jk}^2}{J \times K}$
Kurtosis	$K_{n_i} = \frac{\sum_{j=1}^{J}\sum_{k=1}^{K}\left(\frac{x_{jk}-M(n_i)}{Std(n_i)}\right)^4}{J \times K}$
Mean	$M_{n_i} = \frac{\sum_{j=1}^{J}\sum_{k=1}^{K} x_{jk}}{J \times K}$
Average Energy	$AvgE_{n_i} = \frac{\sum_{j=1}^{J}\sum_{k=1}^{K} \lvert x_{jk} \rvert}{J \times K}$
Standard Deviation	$Std_{n_i} = \sqrt{\frac{\sum_{j=1}^{J}\sum_{k=1}^{K}(x_{jk}-M_{n_i})^2}{J \times K}}$
Entropy	$H_{n_i} = \frac{\sum_{j=1}^{J}\sum_{k=1}^{K}\left(x_{jk}^2 \times log(x_{jk}^2)\right)}{J \times K}$
Skewness	$S_{n_i} = \frac{\sum_{j=1}^{J}\sum_{k=1}^{K}\left(\frac{x_{jk}-M(n_i)}{Std(n_i)}\right)^3}{J \times K}$
Norm	$N_{n_i} = max\left(\sqrt{eig(X_i \times X_i')}\right)$

transform an image to a feature vector. The image is divide into partially overlapping regions/blocks. From these blocks, it generates histograms using the magnitude and orientations of the gradient. The computed histograms are concatenated together to form a final feature vector. More details can be found in [22].

B. Related Work

Mohamed et al. [9] developed an eight-layer convolutional neural network (CNN), improved from AlexNet to classify the mammograms into two classes; scattered density, and heterogeneously dense. They used their own private mammogram dataset with $22,000$ images to effectively train the CNN classifier. Their results demonstrated the effectiveness of utilizing transfer learning compared to training the CNN model from scratch. Trivizakis et al. [3] developed a MBD classification method using pre-trained ImageNet weights for feature extraction, data augmentation on training set, and SVM classifier. They evaluated their method on two mammographic datasets. However, they used overall accuracy as a fitness measure in unbalanced datasets.

Li et al. [7] proposed dilated convolutions and attention modules using ResNet-50 for MBD classification. They used two datasets but their method could not provide good results on the small MBD dataset, since its hard for deep learning models to classify well on limited data. Recently, Valencia et al. [6] proposed a breast density classification method using density map based on texture analysis and fuzzy classification based on subtractive grouping algorithm. Although this method provides a support tool for physicians, the method does not achieve satisfactory classification accuracy on two datasets in the BI-RADS breast density classes.

Zhao et al. [4] developed BASCNet based on the series of ResNet architectures (ResNet-18/34/50/101/152) and adaptive spatial and channel attention network for MBD classification. Two datasets are used to evaluate their method. They employed data augmentation to increase the training data and resizing mammographic images to suit the input size of ResNet architecture. They performed multiple experiments with a series of ResNet-18/34/50/101/152. Though the author performed thorough investigation to increase the classification performance, the method is not interpretable and cannot identify important attributes necessary to discriminate between breast density classes. Li et al. [2] proposed a multi-step process including breast region segmentation, feature extraction using local Binary patterns (LBP) and multi-texture fractal features, feature selection using principal component analysis and an auto-encoder, and MBD classification using SVM. Their method is tested on only one dataset and all images are resized to a single resolution which distorts the aspect-ratio resulting in loss of texture information.

These deep learning methods have several limitations; 1) long training time consuming huge computational resources, 2) usually resize images to suit input configurations of the deep learning model which distorts aspect-ratio and leads to textural information loss, 3) using overall accuracy as fitness measure in unbalanced datasets, and 4) black-box architecture focusing on improving accuracy only and remain hard to interpret prominent attributes.

III. PROPOSED GP METHOD

This study proposes a novel multiple attribute construction using GP (MAC-GP) method for breast density classification using mammographic images. The overall structure of the proposed method is shown in Fig. 3. The test process is shown in Fig. 4. After extracting the statistical features from the original mammographic images and wavelet coefficients, they are concatenated together to form a final feature vector.

A. Training Process

The training data is provided to GP for multiple attribute construction using multi-tree GP. In this work, the number of trees in one GP individual is set to five in order to have enough number of new constructed attributes while keeping the evolutionary process computationally less expensive.

This is the first attribute construction method for mammographic image classification using GP, setting or automatically evolving a suitable number of constructed attributes is still an open issue. Since a GP tree is a mathematical expression, its value can be computed using the original set of attributes. This computed value is the new constructed attribute. Hence,

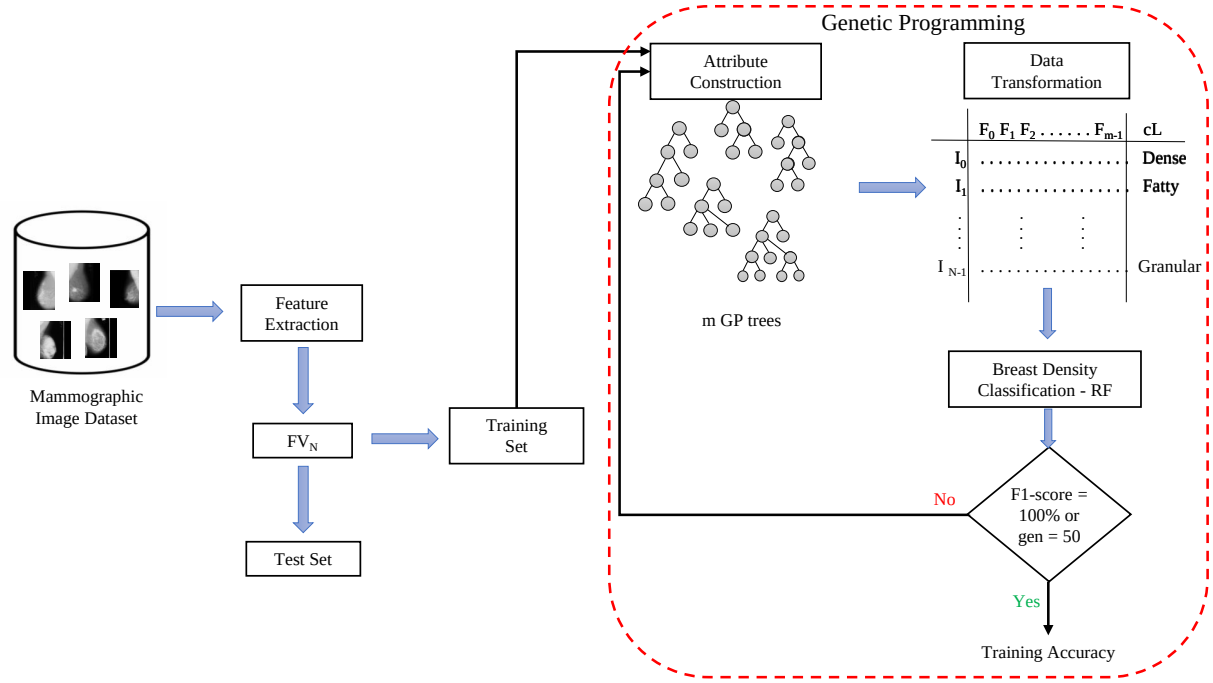


Fig. 3. Training process of the proposed MAC-GP method.

with five trees, five attributes are constructed. This is shown as attribute construction in Fig. 3, where m shows the number of GP trees in one GP individual and the number of constructed attributes. Using these constructed attributes, the original set is transformed to a new training set with m number of attributes. This transformed set is provided to Random Forest (RF) for classification. If the stopping criteria for GP is reached, i.e., a model with 100% F1-score is found or GP reaches 50 number of generations, the RF trained model is used to calculate the classification accuracy. On the contrary, if the stopping criteria is not met, the process is repeated again but this time the GP population includes the highest performing GP individual from the previous generation. Hence, GP explores the search space and keeps generating better or maintain the best attributes so far in the subsequent generations with better discriminating ability between MBD classes.

B. Test Process

After the training process, we get the best GP individual with five trees as shown in Fig. 4. We use these five trees (i.e., constructed attributes) and transform the test data. With this data transformation, we have achieved dimensionality reduction as now the transformed test data at hand has five constructed attributes. This transformed test set is provided to RF for breast density classification which gives the test accuracy and the F1-score.

C. Terminal Set

The terminal set consists of feature vector formed after the feature extraction as discussed in Section II-A. It consists of the following three types of features;

- **Wavelet**: This feature vector consists of a total of 104 features. 8 statistical features are extracted from the original mammographic images using the formulae shown in Table I. 96 features are extracted using three-level tree-structured wavelet decomposition as shown in Fig. 2. (3 levels $\times$ 4 wavelet coefficients $\times$ 8 statistical measures = 96 wavelet decomposition features)
- **LBP**: It consists of 59 features. This study uses only uniform LBP patterns as they can encompass more texture information such as corners, edges, and line-ends to illustrate breast density in mammograms, as compared to non-uniform patterns.
- **HOG**: It consists of 81 features. The HOG histogram is computed from 9 overlapping blocks of a mammographic image where magnitude and orientation information of each block is stored in 9 bins. The bins range from $0°$ to $180°$ where each bin has a range of $20°$. This is adopted from a previous study [23] where it has shown the best performance among the different variants of HOG.

D. Function Set

The function set consists of five operators; one conditional operator $\{if\}$, and four arithmetic operators $\{+, -, \times, /\}$. Among the arithmetic operators, addition, subtraction, and multiplication have their original arithmetic meaning whereas division is protected that returns zero when divided by zero. The conditional operator $\{if\}$ takes three input values and returns the second if the first is greater than the second else it returns the third input value.

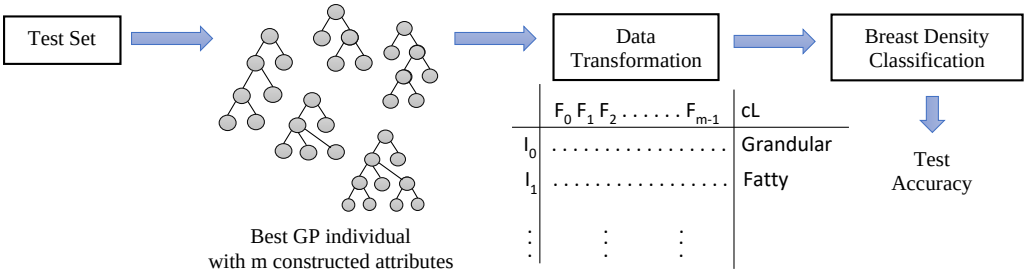


Fig. 4. Test process of the proposed MAC-GP method.

E. Fitness Function

For unbalanced class distributions, balanced accuracy and F1-score are appropriate fitness measures. Balanced accuracygives equal importance in identifying positive and negative cases. However, in medical domains, identifying a positive instance is more important than identifying a negative instance. In breast density classification, identifying a highly-dense MBD image is far more important than identifying a fatty or less-dense MBD image since high density corresponds to developing breast cancer [4]. Considering the importance of correctly classifying the positives, we use F1-score as a fitness function, calculated as:

$$fitness = \frac{1}{z}\sum_{i=1}^{z}\left(2 \times \frac{precision_i \times recall_i}{precision_i + recall_i}\right) \quad (1)$$

$$precision = \frac{TP}{TP + FP} \quad (2)$$

$$recall = \frac{TP}{TP + FN} \quad (3)$$

where z represents the total number of classes, TP represents the true positives, FP represents the false positives, and FN represents the false negatives.

IV. Experiment Design

The datasets are split using 5-fold cross-validation using random stratified sampling. We adopt this setting as the mammographic image datasets used in this study are small (MIAS has 322 images and INbreast has 410).

Being stochastic in nature, GP is executed for 30 times using a different seed value each time. The GP method is implemented using the Evolutionary Computing Java-based package version 23 [24]. In each GP run on each fold, the best individual with the highest performance on the training data having five constructed attributes is used to transform the original set of training data. Hence, the transformed training data consists of five constructed attributes that are used to transform the training data (to train a RF model) and test data (to measure the performance). It is worth mentioning here that the test fold remains unseen during the training process to prevent attribute selection and attribute construction biases.

TABLE II
Parameter Settings of the proposed MAC-GP method.

Parameter	Value	Parameter	Value
Population Size	1024	Tree maximum depth	7
Generations	50	Tree minimum depth	2
Crossover Rate	0.80	Tournament size	7
Mutation Rate	0.19	Initial Population	Ramped half-and-half
Elitism Rate	0.01	Selection type	Tournament
Trees in 1 individual	5	Crossover type	All-index-crossover

A. Benchmark Methods

To evaluate the goodness of our MAC-GP method, we compare it with four commonly used classification methods: NB, SVM with a Radial Basis Function (RBF) kernel, k-NN with $k = 5$, and decision trees (J48), and four ensemble methods: RF with 10 trees, and tree depth of 5, Bagging, Adaboost, and LogitBoost. For implementing these methods, the widely applied Waikato Environment for Knowledge Analysis package version 3.8 [25] is used. In this work, these parameters are empirically defined for MBD classification since they have produced the best results amongst other settings.

In addition, to the best of our knowledge, GP has been studied only once in the past for breast density classification by Burling et al. [18]. We will also look into comparing this method [18] with the proposed MAC-GP method.

B. Parameter Settings

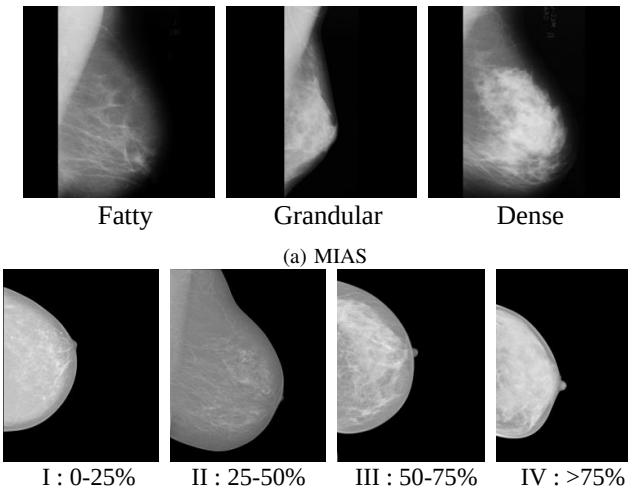
Table II lists the parameter settings of our proposed MAC-GP method. The evolutionary process stops when the RF classifier achieves 100% F1-score or a maximum of 50 generations is reached.

C. Datasets

1) MIAS [26]: The original MIAS Database was digitized at 50 micron pixel edge which was later reduced to 200 micron pixel edge where each image is 1024×1024 pixels. The mini-MIAS[1] is publicly available as a scientific database for research. It consists of 322 mammograms from 161 patients in pgm format.

Expert radiologists have provided image labels in terms of breast density, calcification, architectural distortion, asymmetry, malignancy, and image-coordinates of centre of abnormality. The breast density is categorized into 3 classes:

[1]http://peipa.essex.ac.uk/info/mias.html



(a) MIAS

(b) INbreast

Fig. 5. Image samples from the two datasets where each image belongs to one class.

fatty (106 images), fatty-glandular (104 images) and dense-glandular (112 images). A sample from each of the three classes is shown in Fig. 5(a).

2) INbreast [27]: This database was acquired at the university hospital (Centro Hospitalar de S. Joao Breast Centre) in Portugal. The images were captured from MammoNovation Siemens full-field digital mammograhy, with a pixel size of 70 microns and a contrast resolution of 14-bit. The image size is 3328×4084 or 2560×3328 pixels. There are a total of 410 images which are saved in DICOM format.

Radiologists have categorized the INbreast dataset based on breast density, mass, distortion, asymmetry, lesion annotation, and pectoral muscle annotation. The INbreast database[2] is publicly available for research purposes. The breast density is divided into four categories; $0-25\%$, $25-50\%$, $50-75\%$, and $> 75\%$, as shown in Fig. 5(b).

For the binary classification experiments, we adopted the categorizations by Burling et al. [18] for the two datasets. In MIAS, we combine the fatty and fatty-grandular classes. In case of INbreast, we have two classes with breast density < 50 and > 50.

V. Results and Discussions

Tables III and IV present the results of our experiments of binary and multi-class classification. These Tables are vertically divided into three blocks to show the results of **Wavelet**, **LBP**, and **HOG** features. The values of the MAC-GP results represent the mean and standard deviation among the 30 GP runs, where value of one GP run is computed as the mean of applying 5-fold cross-validation to the datasets. The deterministic methods are run once, hence, their results are represented as the mean of 5-fold cross-validation.

To identify the significance of our MAC-GP method, *One-sample t-test* is applied on the test accuracies. This test checks which method has better discriminative ability to correctly classify the MBD images. Three symbols "+", "−" and "=" are used which represents that the accuracy of MAC-GP significantly outperforms, is significantly worse, and performs similarly in comparison to the corresponding method. The last row in the Tables III, and IV shows the sum of the comparisons between the MAC-GP and the commonly used classification methods based on statistical significance test in terms of "+", "−" and "=" in the respective column.

TABLE III
Results of *binary classification* in terms of balanced accuracy and F1-score using the wavelet, LBP, and HOG feature sets.

		MIAS		INbreast	
	Algorithm	Accuracy	F1-score	Accuracy	F1-score
Wavelet	NB	62.76 +	60.00 +	53.59 −	52.74 =
	SVM	69.63 +	70.01 +	54.00 −	53.51 −
	k-NN	62.50 +	62.46 +	51.72 =	50.72 +
	J48	72.31 −	72.02 −	52.19 −	56.52 −
	RF	70.35 =	70.03 +	52.75 −	56.66 −
	Bagging	70.78 =	71.28 =	50.31 +	53.14 =
	Adaboost	72.04 −	72.22 −	51.32 =	55.48 −
	LogitBoost	69.13 +	70.01 +	51.52 =	50.01 +
	MAC-GP	70.43 ± 1.68	70.75 ± 1.72	51.59 ± 1.20	53.04 ± 1.52
LBP	NB	61.01 +	55.31 +	52.98 +	54.67 +
	SVM	74.57 −	74.21 −	53.57 +	54.96 =
	k-NN	68.35 +	69.26 +	53.74 =	54.22 +
	J48	68.70 +	69.07 +	51.95 +	58.47 −
	RF	65.60 +	66.90 +	52.50 +	53.78 +
	Bagging	70.18 +	71.05 +	52.57 +	57.22 −
	Adaboost	61.89 +	60.34 +	50.65 +	58.60 −
	LogitBoost	71.22 +	71.91 =	51.20 +	57.70 −
	MAC-GP	72.24 ± 1.30	72.65 ± 1.32	54.37 ± 1.12	55.36 ± 1.09
HOG	NB	64.22 +	63.74 +	51.38 +	51.15 +
	SVM	67.93 −	68.08 −	51.73 +	51.70 +
	k-NN	62.17 +	62.42 +	51.68 +	54.26 =
	J48	60.56 +	60.84 +	52.19 +	56.85 −
	RF	66.50 =	67.03 =	51.85 +	52.25 +
	Bagging	61.93 +	62.20 +	53.47 +	53.65 +
	Adaboost	64.21 +	64.29 +	51.82 +	56.25 −
	LogitBoost	63.28 +	63.53 +	52.45 +	55.03 =
	MAC-GP	65.85 ± 1.59	66.05 ± 1.63	54.33 ± 0.90	54.92 ± 1.11
		17+,3=,4-	**17+,3=,4-**	**16+,4=,4-**	**9+,5=,10-**

A. Binary Classification

Table III presents the binary classification results in terms of accuracy and F1-score on the two datasets. MAC-GP results are presented in the last row of each block.

Among the three feature extraction methods, LBP remains prominent in providing better classification performance compared to wavelet and HOG features. On MIAS dataset, MAC-GP achieved on average 70.43%, 72.24%, and 65.85% accuracy using wavelet, LBP, and HOG features. On INbreast dataset, a similar trend is seen where MAC-GP with LBP features provide better classification performance compared to wavelet and HOG features.

For INbreast, it seems that all the four classification algorithms (NB, SVM, k-NN, and J48) remain unable to discriminate dense and sparse tissue images with SVM achieving the highest accuracy at 54.00%. There is a similar trend shown by the ensemble methods on INbreast dataset, where all the four ensemble methods provide accuracies just slightly above 50%. This seems a poor result in case of binary classification.

[2]http://medicalresearch.inescporto.pt/breastresearch/GetINbreastDatabase.html

Though the proposed MAC-GP method provided better classification accuracy, i.e., 54.37% using LBP features on INbreast dataset, it is still not a good binary classification result. Since this is the first attribute construction method using GP, there is a lot of room available to improve the results.

Comparing the results of the eight conventional classification methods with our proposed MAC-GP method, we found that in case of MIAS dataset, SVM remained prominent using the original set of LBP and HOG features. However, on the INbreast dataset MAC-GP outperforms SVM by constructing informative attributes using LBP and HOG features. Hence, it is difficult this way to jump to conclusions as to which classification method is performing better. However, we can consider the overall comparisons made using the statistical significance test. Out of 24 comparisons, MAC-GP significantly outperformed and performed equally well compared to the commonly used classification methods 20 (17+,3=) and 20 (16+,4=) times on MIAS and INbreast datasets, respectively, in terms of balanced classification accuracy (columns 3 and 5). MAC-GP only performed worse 4 times in each of the 24 comparisons on the two datasets. This shows the potential of GP for attribute construction which results in improved classification performance.

B. Multi-class Classification

Table IV presents the results of multi-class classification on the two datasets. Among the three feature extraction methods, i.e., wavelet, LBP, and HOG, LBP features achieved the highest performance with 60.61% average accuracy on MIAS dataset compared to wavelet and HOG features with 58.51% and 47.86% average accuracy, respectively. There is a different trend seen on INbreast dataset where wavelet features provided better classification performance with 45.63% average accuracy as compared to LBP and HOG features which achieved 42.03% and 29.91% average accuracy, respectively.

It is worthwhile to note here that the MIAS dataset has 3 classes with same size of images (1024×1024) in the dataset (easy task) whereas INbreast has 4 classes and has larger and different image sizes (3328×4084 or 2560×3328) (more difficult). The proposed MAC-GP method has provided the best results achieving 60.61% and 45.63% on the MIAS and INbreast datasets, respectively.

From the results of the statistical test presented in Table IV, it is evident that MAC-GP outperforms most of the eight conventional classification methods on the easy (MIAS) and difficult (INbreast) datasets. Out of 24 comparisons, the proposed MAC-GP significantly outperformed and performed equally well compared to the conventional classification methods 22 (17+,5=) and 20 (18+,2=) times on MIAS and INbreast datasets, respectively, in terms of balanced classification accuracy (columns 3 and 5). This shows its effectiveness for MBD image classification problems. MAC-GP only performed worse 2 and 4 times in each of the 24 comparisons on the MIAS and INbreast datasets, respectively.

TABLE IV
RESULTS OF *Multi-class Classification* IN TERMS OF BALANCED ACCURACY AND F1-SCORE USING THE WAVALET, LBP, AND HOG FEATURE SETS.

		MIAS		INbreast	
	Algorithm	Accuracy	F1-score	Accuracy	F1-score
Wavelet	NB	49.70 +	48.57+	39.24 +	27.17+
	SVM	56.20 +	55.34+	35.39 +	35.64+
	k-NN	49.41 +	49.66+	31.28 +	30.15+
	J48	56.40 +	56.31+	43.89 +	44.79+
	RF	55.77 +	55.59 +	38.57 +	38.74+
	Bagging	60.10 −	59.59 −	42.19 +	43.17+
	Adaboost	49.93 +	41.05 +	31.21 +	23.18+
	LogitBoost	57.87 =	57.17+	41.11 +	41.77+
	MAC-GP	58.51 ± 2.34	58.13 ± 2.37	45.63 ± 2.14	46.35 ± 2.10
LBP	NB	43.11 +	40.98 +	36.59 +	28.96 +
	SVM	60.72 =	60.53 =	42.69 =	42.79 =
	k-NN	55.36 +	55.52 +	37.08 +	36.05 +
	J48	51.45 +	51.14 +	33.10 +	33.48 +
	RF	59.24 +	59.08 +	34.73 +	33.00 +
	Bagging	60.12 =	60.22 =	38.33 +	37.34 +
	AdaBoost	47.58 +	38.28 +	26.74 +	18.48 +
	LogitBoost	57.85 +	57.98 +	36.01 +	34.50 +
	MAC-GP	60.61 ± 1.60	60.44 ± 1.67	42.03 ± 1.70	42.20 ± 1.89
HOG	NB	46.57 +	45.81 +	34.07 −	23.47 +
	SVM	55.03 −	54.95 −	31.25 −	31.20 −
	k-NN	48.16 =	47.74 =	25.37 +	23.89 +
	J48	41.22 +	40.49 +	27.73 +	27.25 +
	RF	46.12 +	45.35 +	29.48 =	27.86 +
	Bagging	43.89 +	43.86 +	30.82 −	29.80 =
	AdaBoost	43.42 +	34.05 +	25.54 +	18.00 +
	LogitBoost	47.73 =	47.12 =	32.00 −	31.33 −
	MAC-GP	47.86 ± 1.92	47.12 ± 1.94	29.91 ± 1.47	29.66 ± 1.58
		17+,5=,2-	**18+,4=,2-**	**18+,2=,4-**	**20+,2=,2-**

C. Comparison with the existing GP approach

Burling et al. [18] utilized GP for binary classification of mammographic images. They classified these breast images into fatty and dense tissue classes. They have used 10-fold cross-validation, hence, the proposed MAC-GP method cannot be directly compared. More importantly, they evaluated their method on imbalanced datasets, and used standard overall accuracy as a fitness function which usually leads to bias towards the majority class instances. Hence we remain unable to compare the proposed MAC-GP method with this existing GP approach.

VI. FURTHER ANALYSIS

A. Interpretability of an Evolved GP Individual

GP automatically evolve models that are potentially interpretable. This built-in ability of GP greatly helps to identify the prominent attributes.To show why MAC-GP has achieved good classification performance, we present a good evolved GP individual with five GP trees (constructed attributes) as shown in Fig. 6, achieving 92.53% on the INbreast training data. We analyze the five trees to show the variety of behaviors of attribute selection by GP.

GP selects prominent attributes during the evolutionary process to achieve dimensionality reduction, while still maintaining good classification accuracy. This property of GP is evident from the evolved individual shown in Fig. 6. This GP tree selects 68 unique attributes from a total of 104 attributes while still providing 92.53% classification accuracy. The five

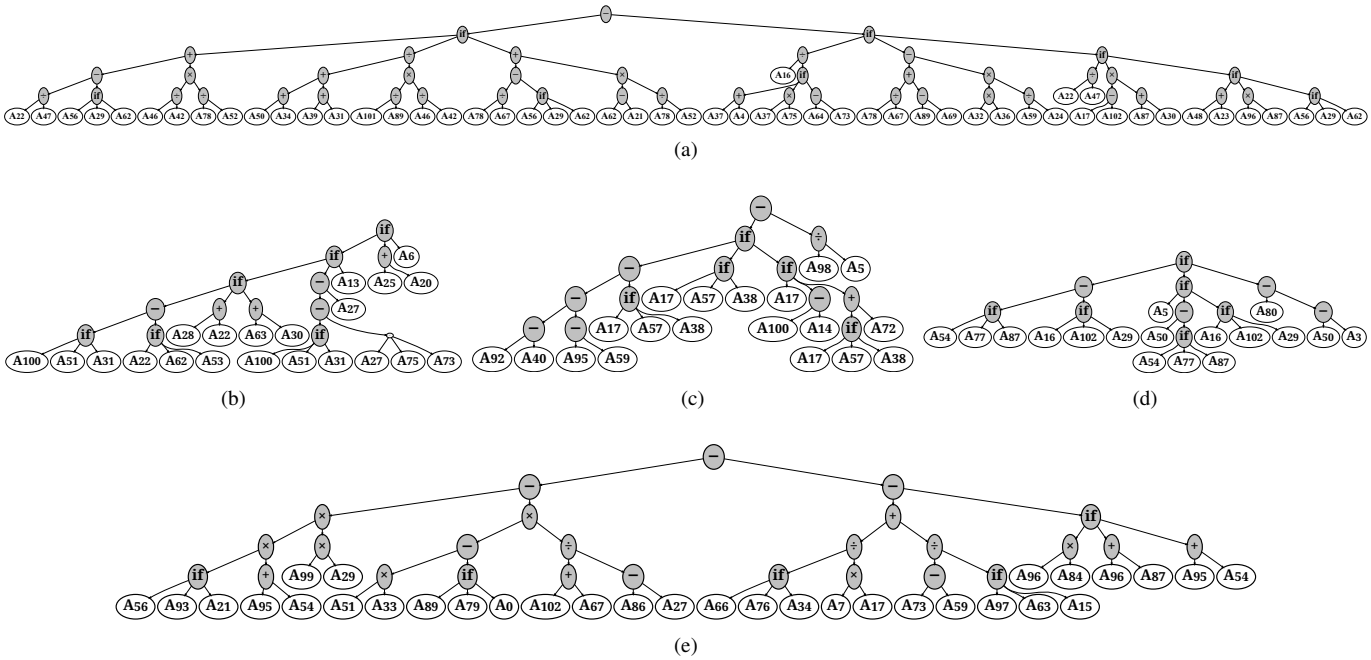


Fig. 6. A good evolved GP individual for $INbreast$ dataset achieved 92.53% accuracy on the unseen data in the multi-class classification task.

trees individually select only 30, 15, 12, 11, and 30 attributes where some attributes overlap among different trees.

Moreover, the sub-trees "*if* (A_{56}, A_{29}, A_{62})", and "*if* (A_{17}, A_{57}, A_{38})" appear three times, which show the importance of these sub-trees to generate this GP individual. For example, the A_{56}, A_{29} and A_{62} attributes are the energy, norm, and entropy, computed from the diagonal coefficients at level-2, level-1, and level-2 decomposition, respectively. The sub-tree "*if* (A_{56}, A_{29}, A_{62})" shows that if the energy value of level-2 diagonal coefficient of a mammographic image is greater than the norm value of level-1 diagonal coefficient, the norm value of level-1 diagonal coefficient is selected, otherwise the entropy value of level-2 diagonal coefficient is selected. For different classes, different values are selected which makes this sub-tree useful in identifying a particular INbreast density class.

A GP individual selects a variety of original attributes to incorporate useful information necessary to discriminate between classes. In the trees shown in Fig. 6, GP has selected attributes from four coefficients (vertical, horizontal, diagonal, and approximation) and three decomposition levels. In addition, it includes all the variety of information provided by the eight statistical measures. Some attributes appeared in this GP individual more than three times, which are listed in Table V. We derive the following observations from this table:

- Among the eight statistical measures, entropy has been selected most often showing its potential for MBD image classification.
- Among the four wavelet coefficients, diagonal and vertical coefficients are picked more often in the evolved GP individual.
- The statistical measures computed from wavelet coefficients appear more often compared to these measures computed from original images. This shows the importance of multi-scale textural image properties of wavelets.

TABLE V
THE DETAILS OF TOP SELECTED ATTRIBUTES IN THE GP INDIVIDUAL SHOWN IN FIG. 6 IN TERMS OF OCCURRENCE, DETAIL COEFFICIENTS, LEVEL OF DECOMPOSITION, AND STATISTICAL MEASURES.

Attribute	Occurrence	Coefficients	Level	Measure
A_{17}	6	Vertical	1	Mean
A_{29}	6	Diagonal	1	Norm
A_{54}	4	Vertical	2	Entropy
A_{56}	4	Diagonal	2	Energy
A_{62}	5	Diagonal	2	Entropy
A_{78}	4	Horizontal	3	Entropy
A_{87}	5	Vertical	3	Average Energy
A_{102}	6	Approximation	3	Entropy

VII. CONCLUSIONS

This study proposes a novel mammographic image classification method where GP automatically constructs multiple attributes for the difficult problem of breast density classification. Different feature extraction methods are utilized to extract texture, shape, local, and global information from mammographic images such as statistical, frequency-based wavelet decomposition, LBP and HOG features. Wavelet decomposition applied on MBD images results in rich wavelet coefficients with discriminative information about the breast density classes. Multi-tree GP has the potential to construct highly informative attributes which help improve the fitness during the evolutionary process.

The results have shown that the proposed MAC-GP method has significantly outperformed and provided similar classifica-

tion performance compared to the eight commonly used classification methods on the two mammographic image datasets. The insights of a good evolved GP individual have shown the trends of selecting prominent attributes with good discriminating ability between different density levels. The results have shown the importance of providing good original attributes with multi-scale textural properties to GP and the good searchability of GP. This is the first attribute construction method for mammographic image classification using GP which has shown that GP has the potential to construct informative attributes and achieve good results for a difficult real-world problem of breast density classification. In addition, setting or automatically evolving a suitable number of constructed attributes is still an open question.

To further improve the classification performance, a preprocessing step before applying feature extraction will be investigated in future. Furthermore, the proposed method in this study does not incorporate domain-specific knowledge that can be utilized in future to improve performance. For example, the domain-specific knowledge provided along with the datasets such as calcification, architectural distortion, asymmetry, and malignancy, and pectoral muscle annotation can be included in the GP individual to look for improvement in the classification performance. This will be employed and investigated in the future.

REFERENCES

[1] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, "Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians*, vol. 71, no. 3, pp. 209–249, 2021.

[2] H. Li, R. Mukundan, and S. Boyd, "Novel texture feature descriptors based on multi-fractal analysis and lbp for classifying breast density in mammograms," *Journal of Imaging*, vol. 7, no. 10, p. 205, 2021.

[3] E. Trivizakis, G. S. Ioannidis, V. D. Melissianos, G. Z. Papadakis, A. Tsatsakis, D. A. Spandidos, and K. Marias, "A novel deep learning architecture outperforming 'off-the-shelf'transfer learning and feature-based methods in the automated assessment of mammographic breast density," *Oncology Reports*, vol. 42, no. 5, pp. 2009–2015, 2019.

[4] W. Zhao, R. Wang, Y. Qi, M. Lou, Y. Wang, Y. Yang, X. Deng, and Y. Ma, "Bascnet: Bilateral adaptive spatial and channel attention network for breast density classification in the mammogram," *Biomedical Signal Processing and Control*, vol. 70, p. 103073, 2021.

[5] C. Balleyguier, S. Ayadi, K. Van Nguyen, D. Vanel, C. Dromain, and R. Sigal, "Birads™ classification in mammography," *European Journal of Radiology*, vol. 61, no. 2, pp. 192–194, 2007.

[6] I. Valencia-Hernandez, H. Peregrina-Barreto, C. Reyes-Garcia, and G. Lopez-Armas, "Density map and fuzzy classification for breast density by using bi-rads," *Computer Methods and Programs in Biomedicine*, vol. 200, p. 105825, 2021.

[7] C. Li, J. Xu, Q. Liu, Y. Zhou, L. Mou, Z. Pu, Y. Xia, H. Zheng, and S. Wang, "Multi-view mammographic density classification by dilated and attention-guided residual learning," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 18, no. 3, pp. 1003–1013, 2020.

[8] Y. Bi, B. Xue, and M. Zhang, "An effective feature learning approach using genetic programming with image descriptors for image classification [research frontier]," *IEEE Computational Intelligence Magazine*, vol. 15, no. 2, pp. 65–77, 2020.

[9] A. A. Mohamed, W. A. Berg, H. Peng, Y. Luo, R. C. Jankowitz, and S. Wu, "A deep learning method for classifying mammographic breast density categories," *Medical Physics*, vol. 45, no. 1, pp. 314–321, 2018.

[10] J. R. Koza *et al.*, *Genetic programming II*. MIT press Cambridge, 1994, vol. 17.

[11] Y. Bi, B. Xue, and M. Zhang, "A divide-and-conquer genetic programming algorithm with ensembles for image classification," *IEEE Transactions on Evolutionary Computation*, vol. 25, no. 6, pp. 1148–1162, 2021.

[12] F. E. Otero, M. Silva, A. A. Freitas, and J. C. Nievola, "Genetic programming for attribute construction in data mining," in *Proceedings of the European Conference on Genetic Programming*. Springer, 2003, pp. 384–393.

[13] D. P. Muni, N. R. Pal, and J. Das, "A novel approach to design classifiers using genetic programming," *IEEE Transactions on Evolutionary Computation*, vol. 8, no. 2, pp. 183–196, 2004.

[14] B. Tran, B. Xue, and M. Zhang, "Genetic programming for multiple-feature construction on high-dimensional classification," *Pattern Recognition*, vol. 93, pp. 404–417, 2019.

[15] T. Singh, N. Kharma, M. Daoud, and R. Ward, "Genetic programming based image segmentation with applications to biomedical object detection," in *Proceedings of the 11th Annual Conference on Genetic and Evolutionary Computation*. ACM, 2009, pp. 1123–1130.

[16] H. Al-Sahaf, M. Zhang, A. Al-Sahaf, and M. Johnston, "Keypoints detection and feature extraction: A dynamic genetic programming approach for evolving rotation-invariant texture image descriptors," *IEEE Transactions on Evolutionary Computation*, vol. 21, no. 6, pp. 825–844, 2017.

[17] Q. U. Ain, H. Al-Sahaf, B. Xue, and M. Zhang, "Automatically diagnosing skin cancers from multimodality images using two-stage genetic programming," *IEEE Transactions on Cybernetics*, 2022, DOI: 10.1109/TCYB.2022.3182474.

[18] F. Burling-Claridge, M. Iqbal, and M. Zhang, "Evolutionary algorithms for classification of mammographie densities using local binary patterns and statistical features," in *Proceedings of 2016 IEEE Congress on Evolutionary Computation*. IEEE, 2016, pp. 3847–3854.

[19] T. Chang and C.-C. J. Kuo, "Texture analysis and classification with tree-structured wavelet transform," *IEEE Transactions on Image Processing*, vol. 2, no. 4, pp. 429–441, 1993.

[20] Q. U. Ain, H. Al-Sahaf, B. Xue, and M. Zhang, "Genetic programming for automatic skin cancer image classification," *Expert Systems with Applications*, vol. 197, p. 116680, 2022, DOI: 10.1016/j.eswa.2022.116680.

[21] T. Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern Recognition*, vol. 29, no. 1, pp. 51–59, 1996.

[22] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, vol. 1. Ieee, 2005, pp. 886–893.

[23] O. L. Junior, D. Delgado, V. Gonçalves, and U. Nunes, "Trainable classifier-fusion schemes: An application to pedestrian detection," in *Proceedings of the 2009 International Conference on Intelligent Transportation Systems*. IEEE, 2009, pp. 1–6.

[24] S. Luke, *Essentials of metaheuristics*, 2nd ed. Lulu, 2013, [Online] Available: http://cs.gmu.edu/ sean/book/metaheuristics/.

[25] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: An update," *Special Interest Group on Knowledge Discovery and Data Mining Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[26] P. Suckling J, "The mammographic image analysis society digital mammogram database," *Digital Mammo*, pp. 375–386, 1994.

[27] I. C. Moreira, I. Amaral, I. Domingues, A. Cardoso, M. J. Cardoso, and J. S. Cardoso, "Inbreast: toward a full-field digital mammographic database," *Academic Radiology*, vol. 19, no. 2, pp. 236–248, 2012.