RESEARCH ARTICLE OPEN ACCESS

An assessment of Diagnosis and Prognosis of Breast Cancer Using Image Mining

Amalendu Bag[1], Manmohan Sahoo[2], Aswini Kumar Mohanty[3]
[1]*Kmbb college of engg*
[2]*Kmbb college of engg*
[3]*The Techno School*

Abstract
The image mining technique deals with the extraction of implicit knowledge and image with data relationship or other patterns not explicitly stored in the images. It is an extension of data mining to image domain. The main objective of this paper is to apply image mining in the domain such as breast mammograms to classify and detect the cancerous tissue. Mammogram image can be classified into normal, benign and malignant class. Experiments have been taken for a data set of 300 images from MIAS dataset of different types with the aim of improving the accuracy by generating minimum no. of rules to cover more patterns. Breast cancer is the top cancer in women worldwide. Early detection of this disease and its classification into cases can improve the prognosis and even save lives by promoting timely clinical management to patients. An accurate diagnosis of breast cancer and its classification into benign, malignant and normal cases is a challenge in cancer research. Because of the ability to enable the computer to learn from past samples to detect and classify patterns, in machine learning, classification algorithms are widely used for cancer identification. However, many of them are focused on binary classification (cancer and no- cancer; benign and malignant). In this work, we present a Computer-Aided Diagnosis (CAD) approach for diagnosis and prognosis of patients into three conditions (malignant, benign and normal) from pixel mammogram images. For the classification task, we explore and compare three outstanding classifiers: Support Vector Machine (SVM), k-Nearest Neighbors (K-NN), and Random Forest (RF) to analyze their accuracy in decision making. In addition, we discuss the effects of pre-processed mammogram images before entering the classifier, which results in higher effective classification.
Keywords: Mammogram, GLCM feature, , Breast Cancer Detection, KNN,SVM, DT Classifiers, Image mining

---

Date of Submission: 28-01-2020 Date Of Acceptance: 13-02-2020

I. INTRODUCTION

Breast cancer is the top cancer in women worldwide which diagnosis, in most cases, is done in the late stages [1]. Early diagnosis can improve the survival possibility by providing timely care, and therefore, it is a crucial health strategy. There is a wide variety of tools and technologies to screen, detect, and diagnose breast cancer. In this sense, mammography is essentially the only widely used imaging modality for breast cancer screening. However, this method is expensive and time consuming. Thus, there is an urge of efficient non-invasive tools, to do so, many approaches were created through the use of mammogram images to serve as a second reader to assist radiologists in the mammographic interpretation process.

The existing approaches mainly work in some common steps, pre-processing tasks, feature ex- traction, and classification. Pre-processing is performed to enhance the mammogram visual quality and perceptibility of the anomalies present in the breasts. Features extraction obtains a set of discriminative and informative data such as texture, abnormality type, mass ratio, color, shape, spatial relations, among others that are used as input to the classification step. Classification allows predicting the class/category of given data features from pixel images.

Several works have begun to explore the effects of machine learning algorithms in many application domains and, particularly, the medical field is one of them. Some algorithms are used to automatically

diagnostic anomalies (as the presence of cancer) present on breast tissues by classify- ing mammogram images and most of them are focused on determining two classes. For instance, in

[2] a fine-tuned SVM demonstrated to be superior to K-NN and probabilistic neural networks (PNN) classifiers to discriminate two breast tumor classes (benign and malignant tumours). To improve their performance, these algorithms have been

combined with signal to noise ratio feature ranking, sequential forward for feature selection and principal component analysis for feature extraction.

In [3] Decision Tree (DT), Multi-Layer Perception (MLP), Naive Bayes (NB), Sequential Mini- mal Optimization (SMO) and Instance-Based for K-Nearest Neighbor (IBKNN) are discussed and combined. Experimental tests performed on three datasets demonstrated that SMO and binary combinations of the evaluated classifiers as MLP with DT and SMO with MLP provide higher accuracy for binary breast cancer classification. K-NN algorithm is studied in [4] for benign and malignant cancer classification to determine the performance of different distances in function of k parameter. The results promote the use of Euclidean and Manhattan distances. Promising results by using RF classifier and feature selection technique are illustrated in [5] for breast mass diagnosis and non-recurrent and recurrent prognostic problem.

An intelligent medical decision model focused on an evolutionary-based strategy for breast cancer detection and recurrence is introduced in [6]. In this approach, some well-performing classifier algorithms as Artificial Neural Networks (ANN), genetic algorithms (GA), SVM, K-NN, and NB compete and collaborate evolutionarily. A K-NN classifier and ANOVA for feature selection are developed in [7]. That approach works in a distributed manner on the scalable Hadoop cluster. An adaptive sparse support vector is proposed in [8]. It classifies microscopic biopsy images into benign and malignant breast tumors by combining the SVM with the weighted L1-norm. SVM with six kernel functions are studied in [9] for increasing reliability for illness breast diagnosis. In [10] suspicious regions from mammogram images are extracted and classified, by using SVM, into three cases: malignant, normal and benign. It deals with pre-processing techniques to improve the image quality and increase the classifier performance.

Innovative automatic methods based on deep learning also have been described for automated mass detection in mammogram images. Deep learning approaches have gained the ability of exten- sively supporting different applications in Artificial Intelligence, Pattern Recognition as well as in Engineering fields due to its significantly precision in patterns recognition and classification from pixel images [11, 12]. In this way, deep learning and RF classifier are explored in [13].

One problem faced by radiologists is that acquired mammogram images often have low qual-ity; having slight dissimilarity between normal, benign, and malignant cancer tissues that leads to inaccurate results. The digitized images took by the mammograph need to be improved, so that image features can be distinguished and they can reflect the subtle variance in the order of many degrees. Thereby, pre-processing tasks for mammogram 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 instance, Histogram Equalization (HE) is used as one of the most popular methods for contrast enhancement, which modifies the gray level histogram of an image to a uniform distribution [14]. But in many cases, it produces over enhancement in the output image and loss of local information. In order to overcome this limitation models such as LCM-CLAHE [15] are proposed. This model conducted an optimal contrast without losing any local information on the mammogram image. LCM-CLAHE consists of two stages of processing to increase the potentiality of contrast enhancement and also to preserve the 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 [16,

II. METHOD

Discovering information from the collected data stored in relational database has been an important work in data mining. The massive collection of image can be mined to discover new information. The main issue of image mining is that it combines the field of content-based retrieval of image, databases and data mining. The image mining process has two phases. The first and the important phase is mining large amount of collected images. The another part is combining the mining of collected data and the corresponding numerical data. Model the image content as a set of blocks, then use any technique to extract the feature. Figure 1 shows the steps in image mining techniques. The collected images are then processed for feature extraction. After extracting the feature from the image mine the content of the image. Evaluate the content of the image with the stored image dataset. The important points to notice on mining images are: Segment images into identifiable regions. Compare the segmented image data with the stored dataset. Apply data mining algorithm to generate object of association rules. An image is accessed once for the feature extraction. The feature extraction results in image blob and image blob descriptors. The image blobe is stored in a file and the blob descriptors are stored in another file. The blob descriptors are used to identify the object of association rules. Once an image is segmented then it is not necessary to search the image content. This work has the objective of classifying mammogram images in three conditions: normal, benign, and

malignant. Its main work flow is shown in Fig. 2. The input mammogram images are enhanced in the pre-processing stage in order to improve its quality and remove undesired information. Then, feature extraction obtains meaningful data to distinguish three different conditions from mammogram image. After, dimensionality reduction techniques are applied, to reduce the amount of discriminative features. Besides, Analysis of Variance (ANOVA) is used to select the most remark- able features from the previous stage. Finally, classification stage is done by using SVM, K-NN and RF supervised classifiers independently, in order to determine their performance on mammogram image classification.

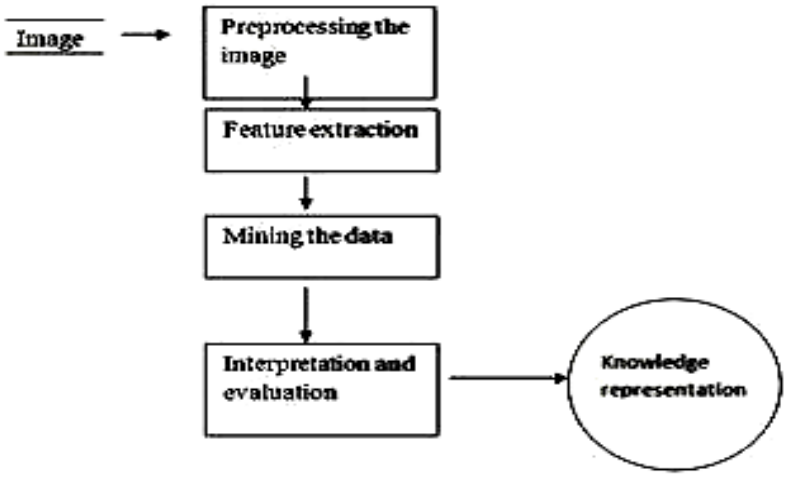

Fig 1.Image mining [24]

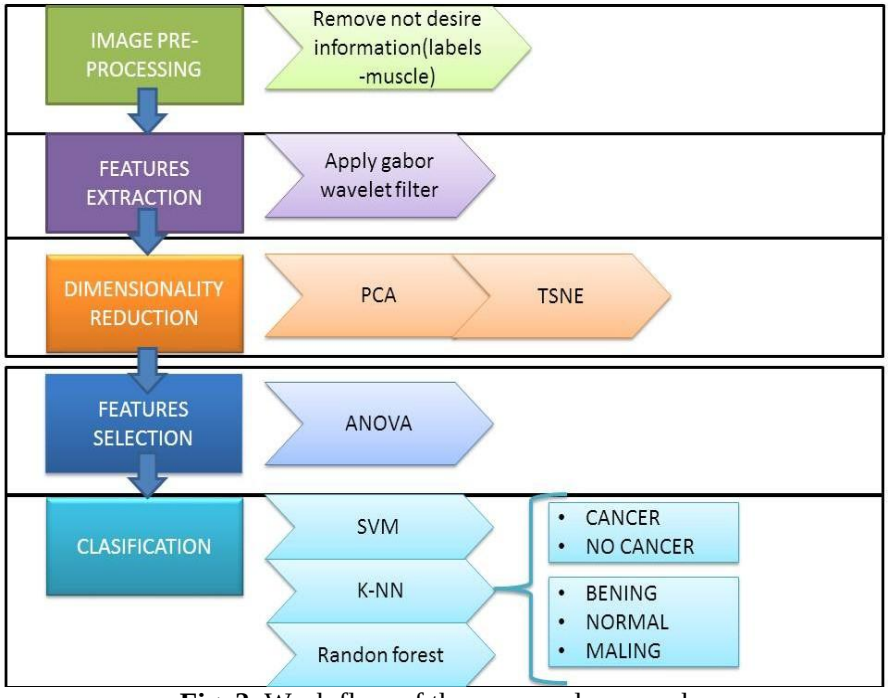

Fig. 2. Work flow of the proposed approach.

Pre-processing: This stage removes undesirable data: labels, margins and pectoral tissues, which can degrade the accuracy of the proposed approach. The pre-processing techniques sequence starts with resizing the images to 1360 796 pixels. Then, a Gaussian filter with a kernel of 55*55 is applied. After, images are binarized using global thresholding $Th1 = 65$. It is followed by Erosion and then dilation with a kernel size of 55* 55. Next, smoothing is applied to images using a kernel size 39 39, and a second binarization is applied through a region of interest using a threshold of $Th2 = 150$ .These parameter values are set for the entire dataset. The thresholding operation is shown in figure 4 of a benign ROI of figure 3.The detailed procedure with steps are shown in figure 8.

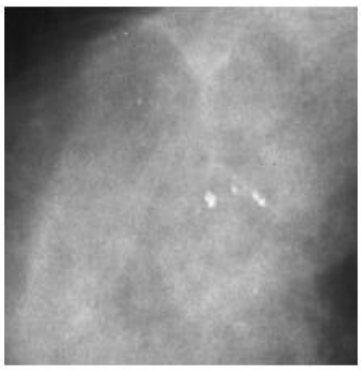

Fig. 3 ROI of a Benign

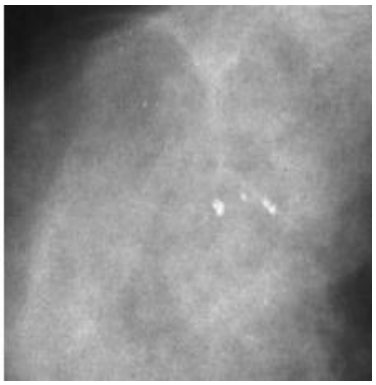

Fig. 4 ROI after Pre-processing threshold Operation

Feature Extraction: It applies Gabor wavelet filters to extract discriminant features from the mammogram images, and therefore, distinguish the three condition types. The mammogram image features are used as a feature matrix for the further process. Firstly, all image pixels are casted into a 32 bit floating point before applying Gabor filter with initial settings based on a kernel size of 33 33, standard deviation of textitSigma = 4, orientation of the normal to the parallel stripes of Gabor function *Theta* = π/16, wavelength of the sinusoidal factor *Lambda* = 10, spatial aspect ratio ***Gamma = 0.5*** and phase offset *PSI = 0*. The different scales and orientations give several patterns as edges, lines spots and flat areas in the mammogram images [19].

Dimensionality Reduction: This stage finds a low-dimensional representation of the feature matrix obtained in the previous stage, because feature extraction often produces massive data, and it is difficult to analyze. Matrix feature has *N M* size, where N is the total number of feature sets and M is the number of mammogram images samples in the dataset. In this approach Principal Components Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (TSNE) are used to reduce the dimensionality of mammogram images based on the work presented in [20]. First, PCA is used saving 95% ratio of variance. Then, TSNE is performed to all data, holding 2 components.

Feature Selection: This stage uses Analysis of Variance (ANOVA) to automatically select, from the output obtained in the previous stage, those relevant features with the most contribution to the prediction variable. It helps in enhancing classifier performance, computational time and cost- effective. ANOVA works mainly for results between groups (class variable) that can be described by variance or whether there is a statistically significant difference between or within groups as shown in table 1.1(a) and 1.1 (b) as well as 1.1(c)

ANOVA algorithms [7] are divided into two parts, the map phase and the reduce phase. In the map phase each mapper reads a row (feature set f_i) from the feature matrix and calculates the required test statistic (F_i) and p-value along with the feature Id (i) as a key-value pair ((i, (F_i, p_i))). It gives this pair into a intermediary file. The reducer then, based on the p-value, decides on whether to select or discard a feature set. It then emits out the selected feature set Ids (((f_{s1}, f_{s2}, f_{s3}, ...))).

Table 1.1(a) Intensity histogram features

Feature Number assigned	Feature
1.	Mean
2.	Variance
3.	Skewness
4.	Kurtosis
5.	Entropy
6.	Energy

Table 1.1.(b) Intensity histogram features and their values

Image Type	Features					
	Mean	**Variance**	**Skewness**	**Kurtosis**	**Entropy**	**Energy**
normal	7.2534	1.6909	-1.4745	7.8097	0.2504	1.5152
malignant	6.8175	4.0981	-1.3672	4.7321	0.1904	1.5555
benign	5.6279	3.1830	-1.4769	4.9638	0.2682	1.5690

Table 1.1©. GLCM Features and values Extracted from Mammogram Image

Feature No	Feature Name	Feature Values
1	Autocorrelation	44.1530
2	Contrast	1.8927
3	Correlation	0.1592
4	Cluster Prominence	37.6933
5	Cluster Shade	4.2662
6	Dissimilarity	0.8877
7	Energy	0.1033
8	Entropy	2.6098
9	Homogeneity	0.6645
10	Maximum probability	0.6411
11	Sum of squares	0.1973,
12	Sum average	44.9329
13	Sum variance	13.2626
14	Sum entropy	133.5676
15	Difference variance	1.8188
16	Difference entropy	1.8927
17	Information measure of correlation	1.2145
18	Inverse difference normalized	0.2863

Classification: The classification process consists of a training phase and a testing phase. In the training phase, the classifier learns from known samples. Meanwhile, the testing phase provides a procedure for the assignment of a class label (mammogram condition type) to the input pattern (mammogram) based on the class labels learned in the training one. This stage uses three ma- chine learning classifiers: Support Vector Machine, K-Nearest Neighbors and Random Forest [20, 21], which are supervised learning algorithms, in order to compare their performance for accurate classification.

Support Vector Machine (SVM): Given labeled training data represented by equation 1, the SVM outputs an optimal hyperplane described by equation 5, which recognizes and categorizes patterns.

$$(x_1, y_1), (x_1, y_1), ..., (x_n, y_n) x_i \in R^d \quad and\ y_i \in (-1, +1) \qquad (1)$$

$$wx^T + b = 0 \qquad (2)$$

Where xi is a feature vector and yi is the class label of a mammogram training i.
The w is the weight vector, x is the input feature vector and b is the bias. w and b satisfy the following inequalities for all elements of the labeled training dataset:

$$wx^T + b \geq 1 \quad if \quad y_i = 1 \qquad (3)$$

$$wx^T + b \leq -1 \quad if \quad y_i = -1 \qquad (4)$$

SVM is mainly focused on solving linearly binary problems by finding w and b so that hyperplane divides input samples into two class with maximum margin $1/\|w\|^2$.
To solve multi-class problems, some kernel functions as linear, polynomial, quadratic, Radial Basis Function (RBF), Gaussian RBF, spline among others can be used to build the kernel method. The kernel method is used to add further dimensions to the input dataset, and thus, make it to a linear problem in the resulting higher dimensional space. A kernel function is represented by:

$$K(x, y) = < f(x), f(y) > \qquad (5)$$

Where K is the kernel function; x, y are n dimensional inputs. f is used to map the input from n to m dimensional spaces. $< x, y >$ defines the dot product. Kernel functions calculate the scalar product between two data points in a higher dimensional space without explicitly calculating the mapping from the input space to the higher dimensional space. Nevertheless, the inner product calculation of two feature vectors often becomes computationally expensive as a result of the feature vector increase in size. The best kernel is often selected through trials because there is no rule to determine which kernel would do the best for a specific pattern recognition problem. Therefore, from the results discussed in [22], this work uses a polynomial kernel function characterized by a low error rate and good classification rate, taking into

account that the selection of kernel function affects notably the performance of the SVM model.
SVM is robust to problems with high dimensional feature space, but not suitable for those that have a huge number of training samples due to its required time for training purposes. In order to illustrate how the SVM algorithm works, its pseudo-code is reported in Fig. 5.

Algorithm 8 SVM algorithm.

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)

```
function RSearchCV(n_neighbors, weights, metrics)
    Create a dictionary with the gamma, kernel, decision function values to try.
    Use Randomized search on hyper parameters
    return best_estimator        ▷ best hyper parameters for knn classifier
procedure SVM_classifier(X_train, y_train, x_test, y_test )
    Use RSearchCV to get the best_estimator
    Fit the model with X_train,y_train
    Predict labels for x_test
    return predicted values    ▷ Score can also be obtained by using y_test
```

Fig 5. The pseudo-code of the SVM algorithm

K-Nearest Neighbors (K-NN): It is based on instances and allows the classification of new elements by calculating their distance to all the other elements *dist*(X_1, X_2). The proper functioning of the algorithm depends on the choice of the distance function used and the value of the parameter *k*, which represents the number of nearby neighbors to the query x_q. The neighbors are weighed by the distance that separates them from the new elements that are classified. This work uses the Euclidean distance discussed in [4], in order to reach a greater precision with a minimum effect due to the variation of the parameter k. K-NN is effective in noisy training data and suitable for cases of a large number of training samples, however, the computation time is increased as much as we need to compute the distance of each instance to all training samples. The main code sequences of K-NN are shown in Fig. 6.

Algorithm 9 k-NN algorithm.

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)

```
function RSearchCV(n_neighbors, weights, metrics)
    Create a dictionary with the n_neighbors, weights, metrics values to try.
    Use Randomized search on hyper parameters
    return best_estimator        ▷ best hyper parameters for knn classifier
procedure KNN_classifier(X_train, y_train, x_test, y_test )
    Use RSearchCV to get the best_estimator
    Fit the model with X_train,y_train
    Predict labels for x_test
    return predicted values    ▷ Score can also be obtained by using y_test
```

Fig. 6. The pseudo-code of the K-NN algorithm.

Random Forest (RF): It is composed of many combined decision trees for classification and regression purposes [5]. It starts selecting random samples from a given dataset. Then. it constructs a decision tree for each sample and gets a prediction result from each decision tree. After, it performs a vote for each predicted result, and finally, it selects the tree with the maximum votes as the pre- diction. RF avoids the overfitting problem by taking the average of all the predictions to cancel out the biases. It can also handle missing values by using median values to replace continuous variables and computing the proximity-weighted average of them. In addition, it promotes the selection of the most contributing features for the classifier. RF is robust for high dimensional spaces and a wide number of training samples. RF pseudo-code is depicted in Fig. 7.

Algorithm 10 Random Forest algorithm.

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)

```
function RSEARCHCV(n_neighbors, weights, metrics)
    Create a dictionary with the criterion, max_features, bootstrap values to try.
    Use Randomized search on hyper parameters
    return best_estimator          ▷ best hyper parameters for knn classifier
procedure KNN_CLASSIFIER(X_train, y_train, x_test, y_test )
    Use RSearchCV to get the best_estimator
    Fit the model with X_train,y_train
    Predict labels for x_test
    return predicted values    ▷ Score can also be obtained by using y_test
```

Fig. 7. The pseudo-code of the RF algorithm.

III. RESULTS AND DISCUSSION

For experimental tests, software routines were implemented in python 3 and WEKA [23][25]. Overall accuracy, precision and specificity defined by equations 6, 7 and 8, respectively, are computed to compare the classification performance reached by each classifier. Accuracy measures the percentage of well-classified mammogram image condition over the total ones. Precision is the proportion of predicted positive conditions that were identified as correct. Specificity measures the proportion of negative conditions that are correctly identified.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (6)$$

$$Precision = \frac{TP}{TP+FP} \quad (7)$$

$$Specificity = \frac{TN}{TN+FP} \quad (8)$$

Where **TN** is the number of correct predictions that confirms the case is negative, **FP** is the number of incorrect predictions that infers a case as positive, **FN** is the number of incorrect of predictions that infers the negative case, and **TP** is the number of correct predictions that confirms the positive case.

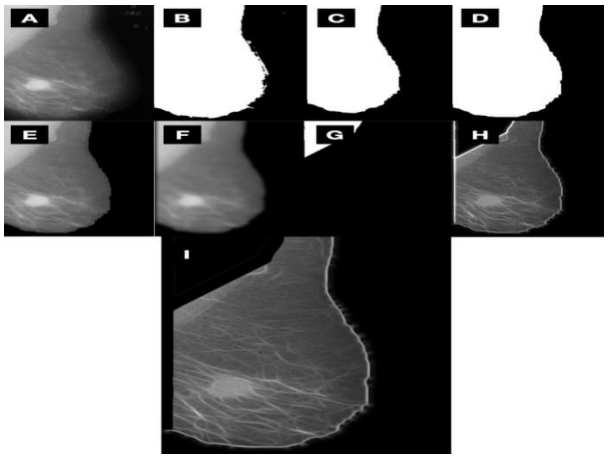

Fig. 8. Steps for Pre-processing images: A) Original image; B) Binarized Image; C) Erode image; D) Dilate image; E) Remove Labels; F) Smooth Image; G) Detect and remove muscle; H) Image applied Gabor wavelet filter; I) Removing not desire edges; J) Image applied original pre-processing method [20].

Above figure 8 illustrates the output sequence obtained from proposed pre-processing stage. It can be seen that the final output presents high contrast and contains only breast tissues. Fig. 8I clearly has a higher image quality in comparison with the raw input image Fig. 8A.

For a fair comparison, all the three classifiers are evaluated for two sequences: with pre-processed mammogram images and raw ones, using 10 forecasts in the cross-validation method, averaged over 10 partitioned times [5]. The average scores presented in Table 2 reports that the pre-processing stage has a crucial impact on the achieved overall accuracy for all evaluated classifiers, so it can be seen an average

increase of 8.8% over the same classifiers that use as the raw mammograms input data. It is apparent from this table that the RF classifier reaches higher performance for classifying two and three mammogram conditions.

Table 2. Overall accuracy achieved by SVM, K-NN and RF classifiers

	3 Conditions			2 Conditions		
Method	**Accuracy**	**Precision**	**Specifity**	**Accuracy**	**Precision**	**Specifity**
SVM	**84.84%**	**73.00%**	**68.00%**	**93.93%**	**46.00%**	**68.00%**
SVM + Preprocessing	**87.87%**	**89.00%**	**89.00%**	**99.00%**	**98.00%**	**95.00%**
k-NN	**86.99%**	**80.00%**	**80.00%**	**98.48%**	**98.48%**	**97.00%**
k-NN + Preprocessing	**89.39%**	**87.00%**	**87.00%**	**99.18%**	**99.00%**	**98.00%**
RF	**84.84%**	**90.00%**	**89.00%**	**98.00%**	**97.00%**	**97.00%**
RF + Preprocessing	**86.36%**	**95.00%**	**94.00%**	**100.00%**	**99.00%**	**98.00%**

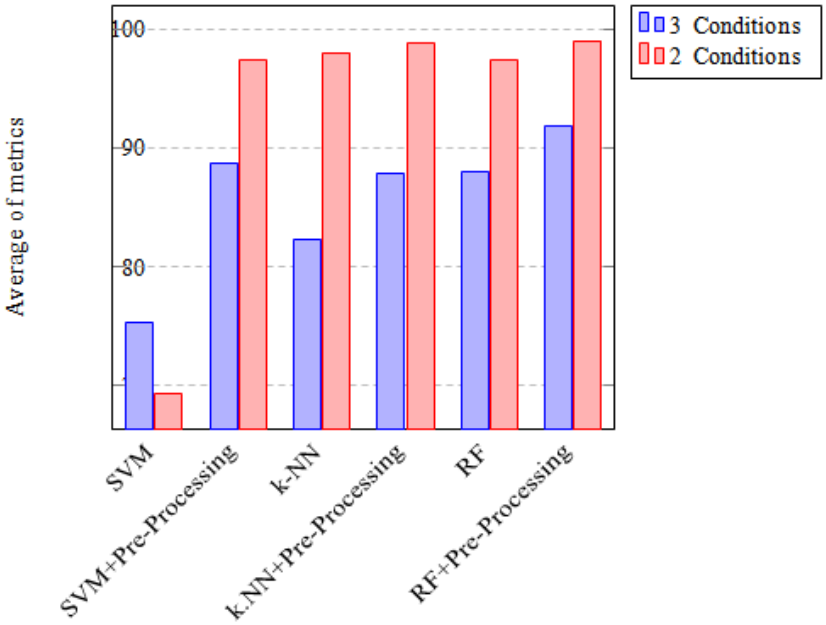

Fig. 9. Average of performance metrics: Accuracy, Precision and Specificity.

On the other hand, the average Accuracy, Precision, and Specificity values depict that almost all classifiers in figure 9 presents a notable reduction in the reached accuracy for multi-class classification in contrast to binary classification, particularly, when it works with pre-processed mammograms, where RF yields the lowest difference and the highest difference is given by k-NN, as shown in Fig. 7. It also supports that the choice of a correct classifier influences enormously on the accurate diagnosis of mammograms.

IV. Conclusion

Mammography is one of the best methods in breast cancer detection, but in some cases radiologists face difficulty in directing the tumors. Further new methods can reduce the computation cost of mammogram image analysis and can be applied to other image analysis applications. The algorithm uses simple statistical techniques in collaboration to develop a novel feature selection technique for medical image analysis. The value of this technique is that it not only tackles the measurement problem but also provides a visualization of the relation among features. In addition to ease of use, these approaches effectively

address the feature redundancy problem. The method proposed has been proven that it is easier and it requires less computing time than existing methods. The proposed approach aims to automate the classification and segmentation process in mammogram analysis. The types of data that need to be classified include normal, benign and malignant conditions. The effects of machine learning algorithms have begun to be explored in several applica- tion domains and the medical field is one of them. In this context, we have tested three supervised predicted models SVM, K-NN and RF to determine their overall accuracy to correctly classify mammogram image conditions. Where RF achieved the highest accuracy metrics for multi-class and binary classification, as well as by using enhanced and raw images. Furthermore, this work has depicted the high impact of image pre-processing sequences for improving accuracy in the classification process.

Focusing on the low classification achieved in the classification for three cases. These results could be improved by: firstly, implementing more data of these conditions, second, engineering new features that lead to the improvement of classifier and finally, changing the parameterization of the pre-processing stage in order to provide higher quality images.

In the future, we would like to develop algorithms that uses simple statistical techniques in collaboration to develop a novel feature selection technique for medical image analysis. The value of this technique is that it not only tackles the measurement problem but also provides a visualization of the relation among features. In addition to ease of use, this approach will effectively address the feature redundancy problem. The method proposed has been proven that it is easier and it requires less computing time than existing methods. We will have a fully automatic system with image mining and classification of the mammogram classes based on the techniques we presented in this paper. It could assist radiologists in the mammographic interpretation process as an appropriate non-invasive tool.

REFERENCES

[1]. Wang, P., Du, Y., Wang, J. (2019). Indentification of breast cancer subtypes sensitive to HCQ-induced autophagy inhibition. Pathology - Research and Practice, 152609. doi:10.1016/j.prp.2019.152609

[2]. Osareh, A., Shadgar, B. (2010) : Machine learning techniques to diagnose breast cancer. In: 2010 5th International Symposium on Health Informatics and Bioinformatics,IEEE, Antalya pp. 114–120.

[3]. Salama, G. I. (2012) : Breast cancer diagnosis on three different datasets using multi-classifiers. Breast Cancer (WDBC) 32(569), 36–43

[4]. Medjahed, S. A. (2013) : Breast cancer diagnosis by using k-nearest neighbor with different distances and classification rules. International Journal of Computer Applications 62(1), 1–5

[5]. Nguyen, C. (2013) : Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic. Journal of Biomedical Science and Engineering 6(5), 551–560

[6]. Gorunescu, F., Belciug, S. (2014). Evolutionary strategy to develop learning-based decision systems. Application to breast cancer and liver fibrosis stadialization. Journal of Biomedical Informatics, 49, 112–118. doi:10.1016/j.jbi.2014.02.001

[7]. Kumar, M., Rath, N. K., Swain, A., Rath, S. K. (2015). Feature Selection and Classification of Mi- croarray Data using MapReduce based ANOVA and K-Nearest Neighbor. Procedia Computer Science, 54, 301–310. doi:10.1016/j.procs.2015.06.035

[8]. Kahya, M. A.(2017) : Classification of breast cancer histopathology images based on adaptive sparse support vector machine. Journal of Applied Mathematics and Bioinformatics 7(1), 49–69

[9]. Wang, H., Zheng, B., Yoon, S. W., Ko, H. S. (2018). A support vector machine-based ensemble algorithm for breast cancer diagnosis. European Journal of Operational Research, 267(2), 687–699. doi:10.1016/j.ejor.2017.12.001

[10]. Azhardeen, M.(2014) : Feature Extraction Based Wavelet Transform in Breast Cancer Diagnosis– A Survey. International Journal of Computer Trends and Technology (IJCTT) **11**(1), 34–37

[11]. Guachi Guachi, Lorena Guachi, Robinson Bini, Fabiano Marinozzi, Franco. (2019). Automatic Col- orectal Segmentation with Convolutional Neural Network. Computer-Aided Design and Applications. 16. 836-845. 10.14733/cadaps.2019.836-845.

[12]. Guachi, L., Guachi, R., Perri, S., Corsonello, P., Bini, F., Marinozzi, F. (2018) : Automatic microstruc- tural classification with convolutional neural network. In: Conference on Information Technologies and Communication of Ecuador, pp. 170–181. Springer, Ecuador

[13]. Dhungel, N., Carneiro, G., Bradley, A. P. (2015) : Automated mass detection in mammograms using cascaded deep learning and random forests. In: 2015 international conference on digital image computing: techniques and applications (DICTA), pp. 1–8. IEEE, Australia

[14]. Makandar, A. (2016) : Pre-processing of mammography image for early detection of breast cancer. International Journal of Computer Applications **144**(3), 11–15

[15]. Muneeswaran, V., Rajasekaran, M. P. (2019) : Local contrast regularized contrast limited adaptive histogram equalization using tree seed algorithm—An aid for mammogram images enhancement. In: Smart Intelligent Computing and Applications, pp. 693–701. Springer, Singapore

[16]. Robert, K. (1981) : Cubic convolution interpolation for digital image processing. IEEE transactions on acoustics, speech, and signal processing **29**(6), 1153–1160

[17]. Parker, J. A., Kenyon, R. V., Troxel, D. E. (1983). Comparison of Interpolating Methods for Image Resampling. IEEE Transactions on Medical Imaging, 2(1), 31–39. doi:10.1109/tmi.1983.4307610

[18]. Das, S. (2014) : Medical image enhancement techniques by bottom hat and median filtering. Interna- tional Journal of Electronics Communication and Computer Engineering **5**(4), 347–351

[19]. Lladó, X., Oliver, A., Freixenet, J., Martí, R., Martí, J. (2009). A textural approach for mass false positive reduction in mammography. Computerized Medical Imaging and Graphics, 33(6), 415–422. doi:10.1016/j.compmedimag.2009.03.007

[20]. Raghavendra, U., Rajendra Acharya, U., Fujita, H., Gudigar, A., Tan, J. H., Chokkadi, S. (2016). Application of Gabor wavelet and Locality Sensitive Discriminant Analysis for automated identifi- cation of breast cancer using digitized mammogram images. Applied Soft Computing, 46, 151–161. doi:10.1016/j.asoc.2016.04.036

[21]. Han, J., Pei, J., Kamber, M. (2011) : Data mining: concepts and techniques. 3rd edn. Elsevier, USA

[22]. Sangeetha, R., Kalpana, B. (2010). A Comparative Study and Choice of an Appropriate Kernel for Sup- port Vector Machines. Communications in Computer and Information Science, 549–553. doi:10.1007/978- 3-642-15766-093

[23]. Chachalo B. et al. (2019) Automated Identification of Breast Cancer Using Digitized Mammogram Images.In: Nystr¨om I., Hernandez Heredia Y., Mili´an Nu´n˜ez V. (eds) Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. CIARP 2019. Lecture Notes in Computer Science, vol 11896. Springer, Cham

[24]. A. A. Khodaskar; S. A. Ladhake, "Image Mining: An Overview of Current Research" , Fourth International Conference on Communication Systems and Network Technologies, 2014.

[25]. www.cs.waikato.ac.nz