Information Theory-based Detection of Noisy Bit Planes in Medical Images

Hugo N. de Oliveira, Jefersson A. dos Santos
Departamento de Ciência da Computação
Universidade Federal de Minas Gerais, UFMG
Belo Horizonte, Brazil
Email: {oliveirahugo,jefersson}@dcc.ufmg.br

Matheus C. de Melo, Thaís G. do Rêgo, Leonardo V. Batista
Centro de Informática
Universidade Federal da Paraíba, UFPB
João Pessoa, Brazil
Email: {matheus.melo,gaudencio.thais,leonardo}@ci.ufpb.br

***Abstract*—Mammographic Computer-Aided Diagnosis systems are applications designed to assist radiologists in diagnosis of malignancy in mammographic findings. Most methods described in the literature do not perform a proper preprocessing step in mammographic images prior to classification, which can generate inconsistent results due to the potentially large amount of noise in medical images. This paper proposes a new method based on Information Theory and Data Compression for detection of random noise in image bit planes. In order to validate the efficiency of the proposed noise removal method, we used Machine Learning algorithms to classify mammographic findings from the Digital Database for Screening Mammography. Results using texture features indicate that a reduction in the radiometric resolution of 4 or 5 bit planes in digitized screen film mammographic images result in a better classification performance.**

***Keywords*-noise detection; mammogram classification; information theory; data compression**

I. Introduction

Breast cancer is the most common type of cancer among women and it is responsible for one of the largest mortality rates, second only to lung cancer [1]. Tan *et al*. [2] estimates that one in eight women will develop breast cancer in her life time. According to Li, Williams and Bottema [3], generally the incidence rates of breast cancer are higher in developed countries than in developing countries, but the incidence has increased in most regions over the last few decades, mainly in developing countries. Although breast cancer has a good prognosis, high death rates are mainly related to late diagnosis [4]. Thus, research in development and improvement of detection and diagnosis methods for breast cancer on early stages is essential.

Clinical breast exam, ultrasound imaging, magnetic resonance and mammography are some of the detection methods used to detect breast cancer. Currently, mammography is the most effective tool for early detection of breast cancer [5]. A screening mammogram is a grayscale image and usually has 12 to 16 bits per pixel (b/p) of radiometric resolution. This large number of bits per pixel is required to represent all the information captured in the exam and facilitate the identification of possible nodules. Analysis of mammograms is an error-prone task, as factors like fatigue, distraction and poor experience of the radiologist can lead to unnecessary biopsies or wrong diagnosis [6, 7], as only about 20% to 50% of patients referred for biopsy are found to have a malignancy [8].

Since the 1970's, many studies [9, 10, 11] have assessed through statistic evaluation the efficiency of mammographic screening in preventing breast cancer mortality. According to Berry *et al.* [10], the decrease of 24% in mortality rate of breast cancer patients observed between 1990 and 2000 is likely explained by early detection techniques based on mammographic imaging and advances in treatment.

Computer-Aided Diagnosis (CAD) systems are applications that assist radiologists in classification of possible nodules or others abnormal findings. Generally, mammography CAD systems are developed with the objective of exposing relevant findings in mammograms, extract useful diagnostic characteristics and apply Machine Learning algorithms to classify the findings, like malignant and benign nodules or microcalcifications. For CADs to achieve good results, besides obtaining relevant descriptors from image sets, they should execute proper preprocessing steps in biomedical images. Jalalian *et al*. [5] defines the preprocessing step as one of main stages in CAD system.

In order to obtain acceptable performances and be useful to physicians, CAD systems should be validated with different databases. The mammographic image database most frequently used in literature is the Digital Database for Screening Mammography (DDSM) [12] due to its vast amount of images spanning several scanners, complete annotations, Region of Interest (ROI) delimitation for each finding, presence of anatomopatological data and to the fact that it is publicly available on the Internet. Screen films were used in the original acquisition process for DDSM images, which were later digitized. The acquisition process for Full Field Digital Mammography (FFDM) – which comprises the current state-of-the-art in mammographic imaging – consists of mapping intensities detected by photosensitive cells directly to electric impulses without any analogical middle step. This process potentially reduces the amount of noise generated in the digitization of x-ray data originated from these mammograms.

Many papers describing methods for detection and diagnosis of breast nodules use the DDSM but do not perform any preprocessing step or simply apply a scalar quantization on the data without proper statistical evaluation. Dhungel, Carneiro

and Bradley [13] use the DDSM but do not perform any preprocessing technique on the data. Li, Williams and Bottema [3] describe their preprocessing steps but do use any noise removal method. Chen, Lan and Ren [14] developed and tested a CAD system with images from the DDSM, however proper preprocessing was not performed. Rehman, Chouhan and Khan [15] used contrast enhancement and applied a mean filter with a 5x5 mask, which can insert artificial artifacts in images, hampering analysis through the use of texture features. Vieira *et al*. [16] proposed a denoising algorithm for quantum noise reduction in digital mammography using Wiener filtering.

The main contribution of this paper is the description and validation of a method for detection and removal of noise in biomedical images. Our methodology was validated as a preprocessing step for automatic malignancy classification of breast nodules in mammograms using texture features. The proposed technique is based on Information Theory and Data Compression and was tested for the classification of masses extracted from images in the DDSM dataset. The presented low cost methodology may be used as a preprocessing step for other mammographic CAD systems based on texture features and should be extensible for other areas of biomedical image processing and general purpose image processing. The other sections of this paper are organized as follows: Section II describes the proposed methodology for noise estimation; Section III introduces the classification method comprising the texture features and Machine Learning algorithms used in our analysis; Section IV shows the empirical results originated from our test procedure and discusses them; and, at last, Section V presents our conclusions about the results.

II. Information Theory-based Noise Estimation

The high radiometric resolution in mammograms – usually ranging between 12 and 16 b/p – may contain Bit Planes (BPs) with useless information, potentially hampering the extraction of texture features from these images. Random pixel intensity variations introduced by the noisy channel may be mistaken as the inherent texture variabilities of mammographic images, which can lead to errors in malignancy classification. While BPs near the Most Significant Bit Plane (MSPB) show highly correlated samples and complex patterns, BPs closer to the Least Significant Bit Plane (LSBP) exhibit mostly randomness. Visual analysis of digitized mammographic images often reveal the large amount of noise in the LSBPs of those images, as show in Fig. 1, which only starts to show recognizable patterns after Fig. 1(f).

A. Information Theory

In order to isolate noise in mammographic images, a new methodology based on Information Theory for BP random noise detection was developed. While visual inspection is useful to roughly detect noisy BPs, a more thorough statistical analysis is necessary to correctly assess the amount of BPs corrupted by noise. Information Theory [17] provides an objective measure of the amount of randomness in a message

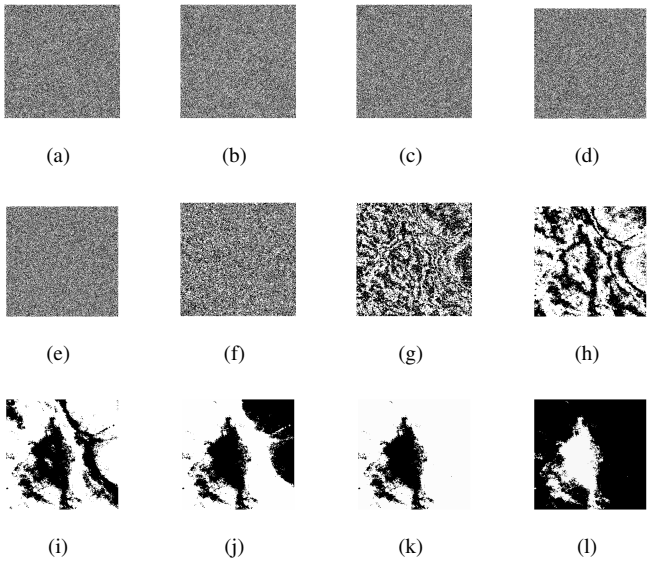

Fig. 1. BPs of mammographic ROI extracted from the DDSM. (a) LSBP. (b)-(k) Intermediary BPs. (l) MSBP.

emitted by a given source. The information I_{x_i} of a symbol x_i represented in the base b with probability $\mathbb{P}(X_i = x_i) = p_{x_i}$ is given by:

$$I_{x_i} = \log_b (p_{x_i}) \quad b\text{-based digits.} \tag{1}$$

Information may be understood as the number of b-based symbols necessary to encode x_i. As computers use binary encoding, information is often expressed in bits, therefore Eq. 1 for binary digits is rewritten as:

$$I_{x_i} = \log_2 (p_{x_i}) \quad \text{bits.} \tag{2}$$

The entropy $\mathbb{H}(X)$ of a message X is defined based on the probabilities $p_{x_1}, p_{x_2}, ..., p_{x_n}$ for the occurrence of the symbols $x_1, x_2, ..., x_n$ and their respective values of I_{x_i}. The calculation of the entropy $\mathbb{H}$ for the message X is given by:

$$\mathbb{H} = -\sum_{i=1}^{n} p_{x_i} I_{x_i} = -\sum_{i=1}^{n} p_{x_i} \log_2 (p_{x_i}) \text{ bits/symbol.} \tag{3}$$

Alternatively to the entropy, one can understand the randomness of a message in terms of its Compression Ratio (CR) after compacted by a compression algorithm. Performing an analysis using CR is equivalent to evaluating the entropy of a message, as the compression performance of a lossless data compression algorithm is bounded by the message's entropy. CR is calculated as follows:

$$\text{CR} = \frac{M}{m}, \tag{4}$$

where M represents the size (in bits) of the original message and m is the size after compression. Higher values of $\mathbb{H}$ result in lower values of CR. A CR of 1:1 means that the message could not be compressed any further, implying that the original message was already encoded using the least amount of bits

allowed by its entropy. Therefore, for a message composed of a string of random symbols, CR $\approx$ 1:1.

In order to compare CRs, one should take into account the performance of the compression algorithm when fed a string of random values of the same size, as most state-of-the-art compression schemes achieve slightly suboptimal codes. One should expect a significant statistical difference between average CRs of a randomly generated sample of binary values and an actual set of binary samples obtained from a BP extracted from an image.

B. Hypothesis Testing

According to the definitions presented in Section II-A, it is possible, then, to define a methodology based on statistical hypothesis testing to evaluate differences between CRs:

For $\alpha = 0.0025$ and two paired samples A and B of CRs, where A was obtained from the compression of a set of BPs from mammographic images and B was calculated after compression of random strings of 0's and 1's, the following one-tailed hypothesis test for paired difference of mean values is proposed:

$$H_0 : \mu_A = \mu_B$$
$$H_1 : \mu_A > \mu_B.$$

In this setup, the confidence level (p-value) of the hypothesis test is $p = 99.75\%$ and the limit for the z-score in order to reject H_0 is the value 2.8070. The p-value must be empirically tested in order to adapt to the amount of noise in the dataset.

While it is not correct to assume similar distributions for A and B when H_0 is not refuted, it is possible to derive a methodology for assessing noisy BPs using this approach. When μ_A is not significantly larger than μ_B, the BPs that generated A will be considered highly infected by noise and statistically indistinguishable from random noise. In contrast, when H_0 is refuted, the BPs that originated A will be identified as containing useful information. Due to the large number of samples analysed in our tests, we used the z-test to validate the hypothesis that the CRs from noisy BPs are statistically indistinguishable from CRs obtained from the compression of random noise.

C. Compression Algorithm

In order to analyse noise using CR, one should choose a state-of-the-art compression algorithm, such as the Prediction by Partial Matching (PPM) [18] scheme. The PPM algorithm uses previous symbols read in the message to predict the current symbol using a contextual model controlled by the parameter K, which represents the highest context the PPM uses to encode symbols. Let the message X be composed of the values $x_1, x_2, ..., x_n$ and each symbol x_i be written using an alphabet comprised of the words $w_1, w_2, ..., w_m$. If a PPM is executed with $K = 1$, for each i^{th} sample X_i, instead of encoding the probability $\mathbb{P}(x_i = w_a)$, the PPM will encode the conditional probability $\mathbb{P}(x_i = w_a|K_1)$, where $K_1 = \{x_{i-1} = w_b\}$. This encoding shortens the number of symbols considered in the calculation of the entropy $\mathbb{H}$. If the current context $K_{i-1} = w_b$ was not encoded previously, the PPM encodes it using $K = 0$. For $K = 2$ the probability encoded is $\mathbb{P}(x_i = w_a|K_2)$ if the context $K_2 = \{x_{i-1} = w_b \cap x_{i-2} = w_c\}$ was previously encoded. If K_2 was not seen before, the PPM tries to encode x_i using the context K_1.

The Binary Prediction by Partial Matching (BPPM) algorithm – first described by Marques *et al.* [19] – simplifies the data structure used to represent the contextual probabilities of the classic PPM, allowing for an implementation using only a small fraction of the memory cells needed by the original algorithm. The static probability representation adopted by the scheme also shortens the execution time for the codification of each symbol, as the access to the probabilities becomes random – that is, BPPM allows for direct access to the memory position of the each symbol's probability rather than a search procedure, as performed by the classic PPM. This simple implementation is possible because, while the PPM encodes an input symbol as a whole, the BPPM encodes it one bit at a time. Fig. 2 shows an example of the codification of a message M containing the samples $\{15, 10, 7, 8, 5, 1, 0\}$, each one with 4 BPs.

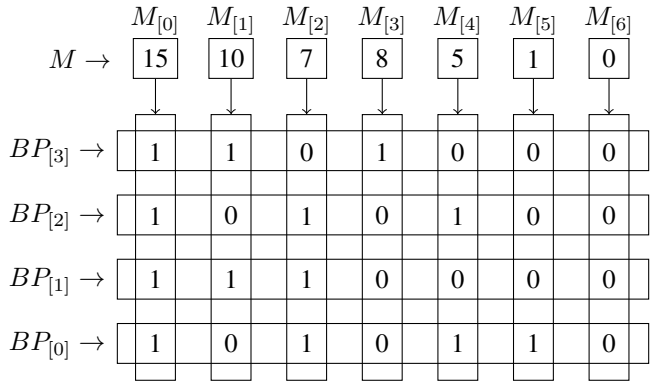

Fig. 2. BP coding sequence of a BPPM.

In Fig. 2, the first bit analysed is the one in row $BP_{[3]}$ and column $M_{[0]}$, which feeds the probabilities in the contextual modelling of the MSBP. The second bit analyzed by the compressor is in row $BP_{[2]}$ and column $M_{[0]}$, using the table of probabilities computed for this BP. The third and fourth encoded bits are the ones in column $M_{[0]}$ and rows $BP_{[1]}$ and $BP_{[0]}$ respectively. After all BPs of $M_{[0]}$ are encoded, $M_{[1]}$ is passed to the BPPM, which processes the sample in the same order from $M_{[0]}$. Other samples are encoded sequentially until the LSBP ($BP_{[0]}$) of symbol $M_{[6]}$ is compressed.

As pointed by Brasileiro and Cavalcanti [20], a hardware implementation of the PPM algorithm requires a preallocation of all memory blocks for each symbol in the alphabet and all possible contexts. Therefore, the number of memory cells T_P needed for the execution of a classic PPM scales exponentially, according to:

$$T_P = B \sum_{k=0}^{K} \left(2^W\right)^{(k+1)} = B \sum_{k=0}^{K} 2^{W(k+1)}, \qquad (5)$$

where W is the amount of bits used to represent each word, B is the number of bytes used to represent integers in the algorithm and K is the context size.

The BPPM scheme operates independently over the BPs of the image, that is, ignoring correlations between BPs while taking into account only the dependency between bits of the same BP in neighbor samples. For each $BP_{[n]}$, with $0 \leq n < W$, the BPPM creates a contextual probabilistic model. For a context size K, each BP needs a number of memory cells equal to twice the amount of possible contexts, one for the probability of the bit 0 and the other for the bit 1. Therefore the expression for the amount of probabilities stored by each BP is 2×2^K, or 2^{K+1}, which may each one require an amount of B bytes to represent. The total number of cells T_B used to represent all probabilities in the contextual modelling of a BPPM is equal to:

$$T_B = 2^{K+1} W B. \tag{6}$$

In Eq. 6, the word size W is a multiplication constant, while in Eq. 5 it contributes exponentially to T_P. The classic PPM also requires memory cells for the modelling of all context sizes (from 0 to K), while the BPPM only needs one level of contextual modelling with size K. For instance, using the parameters $K = 2$, $B = 1$ and $W = 16$ ($2^{16} = 65536$ possible symbols), $T_P = 2.815 \times 10^{14}$ and $T_B = 128$. This drop in memory requirements allows for hardware implementations of the BPPM.

Prediction-based compression methods like the PPM and BPPM schemes provide, in addition to CRs close to the message's entropy, the possibility of CR analysis using different contextual levels by changing the parameter K. This contextual analysis allows for a more robust prediction, which grants the schemes the ability to predict samples based on more complex patterns. In other words, even if visual inspection of BPs cannot detect complex contextual patterns hidden in the binary image, they should be detectable by analysing the CRs of a BPPM with large values of K and comparing them to the CRs of a randomly generated sample.

III. Experimental Setup

The classification procedure used in our tests is schematized in Fig. 3. Firstly we extracted 160 randomly selected mammograms from DDSM [21]. DDSM is composed of about 2,600 cases, each one containing two images from each breast: a Craniocaudal (CC) and a Mediolateral Oblique (MLO) view. Mammograms in this database are encoded as grayscale images with 16 b/p, but most of them contain only 12 useful bits of information. In addition to showing the diagnosis confirmed by pathological and anatomical exam, DDSM also contains metadata about all patients.

Mammographic ROIs are findings considered relevant by a physician for the diagnosis of breast cancer. Asymmetries, architectural distortions, masses, stroma distortions and microcalcifications are the most common types of mammographic findings. This work will be focused on the study of masses due to the fact that they are one of the most common findings

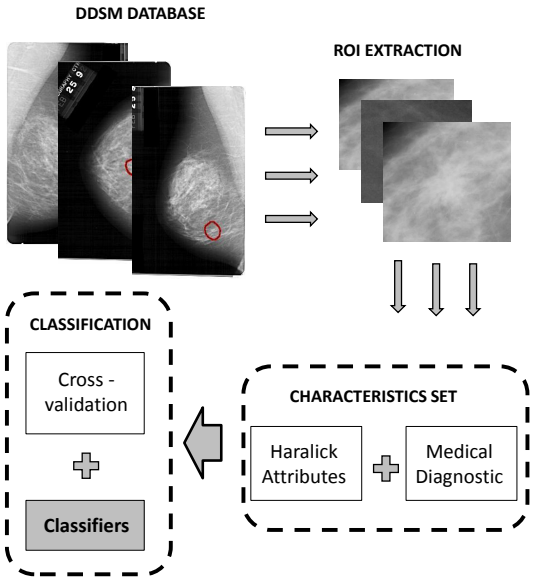

Fig. 3. Illustration of the sequence of steps for classification: ROI delimitation, extraction of feature sets from ROIs and classification.

seen on mammography [12]. After the acquisition of the full images, ROIs were extracted from the mammograms using the ground-truths provided by the DDSM.

Automatic classification of mammographic findings in malignant or benign is a difficult procedure due to the low contrast of images, the presence of noise and due to the distortions inserted by microcalcifications and masses. Thus, it is important to use texture characteristics for methods of automatic detection and classification [11]. Many methodologies [11, 22, 23, 24] include Haralick descriptors [25] in their feature sets for detection and classification of mammographic findings. Therefore, in our work 13 Haralick texture features were extracted from each ROI. Prior to our noise detection/removal method, the subset of images selected from DDSM had initially 12 b/p. In order to analyse the influence of noise in classification, Haralick features were extracted from images after each consecutive BP removal. Thus, texture characteristics are extracted from images with 12 b/p, 11 b/p, 10 b/p and so on until 4 b/p, resulting in 9 feature sets.

As presented in Fig. 3, the classification stage used Machine Learning algorithms to classify between malignant and benign images with cross-validation due to the relatively small number of images. Following the methodology adopted by related works, like Acharya *et al.* [26], 10 folds were used. This work uses the following classifiers implemented in the Sklearn framework [27]: Naive Bayes (NB) [28], Support Vector Machine (SVM) [29], Decision Tree (DT) [30], Random Forest (RF) [13] and a variation of Decision Trees, the Extremely Randomized Trees (ET) algorithm [31].

Kernel functions can be specified for the SVM algorithm in order to transform the data and allow for a linear separation using an optimal hyperplane. The results in this paper were

achieved with linear kernel. The RF and ET algorithms have the $n\_estimators$ parameter that defines the number of trees in their forests. This parameter was set to $n\_estimators = 10$. The accuracies achieved using each classifier will be presented in Section IV.

IV. Results

Table I presents the z-scores for the hypothesis test described in Section II for different BPPM contextual levels and for the 8 LSBPs of DDSM mammograms. Green values indicate the cases where H_1 was accepted and H_0 was rejected, while red values show the results where H_0 could not be refuted. Means from the CRs of MSBPs present significant differences from means of CRs computed from the randomly generated BPs, while tests using 4 or 5 of the LSBPs cannot refute H_0. For these tests, only 141 images from our sample of 160 images were used, as some ROIs are located close to the borders of the mammogram and, therefore, contain background, which would compromise the results of the hypothesis testing by artificially enlarging the CRs' average values. In order to fill 16 bits (2 bytes), DDSM images also present an unusual encoding wherein neighbor pixel intensities are not encoded as neighbor integers, as shown in [32]. Exploratory tests revealed the need for a mapping from this uncommon encoding to the 4096 intensities of the 12 usable b/p in those images, that is, the interval $[0..4095]$. The values did not completely fill the interval $[0..4095]$ because not all the $2^{12} = 4096$ possible pixel values were present in the original DDSM images. The outlier CR values in the BP 2 are likely explained due to this histogram rescaling process, which introduced statistical sampling artifacts into the BPs. The p-value for the z-scores was empirically set to 99.75% in order to ignore these outliers.

TABLE I
z VALUES FOR THE BPS OF 141 IMAGES FROM THE DDSM DATABASE USING THE CONTEXTS $K = 1, 2, 3, 4, 5, 10, 15$.

	k_1	k_2	k_3	k_4	k_5	k_{10}	k_{15}
BP 1	0.54	0.40	0.16	0.16	0.08	0.01	-0.02
BP 2	2.72	1.71	0.97	0.61	0.33	0.05	0.08
BP 3	0.98	0.64	0.44	0.33	0.15	0.04	0.04
BP 4	1.77	1.01	0.61	0.37	0.17	0.05	-0.03
BP 5	3.22	3.01	2.55	1.92	1.36	0.20	0.19
BP 6	6.25	6.23	6.23	6.22	6.21	5.31	5.67
BP 7	11.51	11.65	11.74	11.77	11.78	11.48	11.08
BP 8	21.74	23.34	24.05	24.20	24.22	23.73	22.49

Visual analysis of Table I reveals the large gap between the values of CRs for BPs 4 and 6. While the first four BPs present z-values close to 0 – mainly for higher values of K – the fifth BP already had two rejections of H_0 for $K = 1$ and $K = 2$. This result means that, while BP 5 does not yet present complex structures, it already contains an imbalance in the distribution of 0's and 1's in most contexts of size 1 and 2, therefore CRs for $K = 1$ and $K = 2$ are significantly larger than CRs generated randomly.

We performed two kinds of validations for our estimation of noisy BPs previously to the extraction of texture features. The first one was an analysis of average noise using the estimation proposed in [33, 34], presented in Fig. 4. Visual inspection of Fig. 4(a) indicate the nearly linear growth of noise when using less than 9 BPs, while Fig. 4(b) shows the exponential nature of noise growth considering 9 or more BPs.

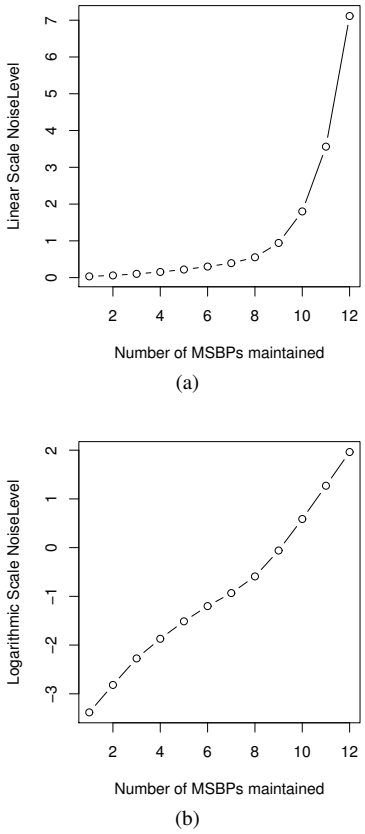

Fig. 4. NoiseLevels obtained from the algorithm described by Liu *et al.* [33, 34]. (a) Linear scale. (b) Logarithmic scale.

The results shown in Table I match the estimation of Fig. 4, with the benefit of having a much lower computational cost, as the estimation technique introduced by Liu *et al.* [33, 34] is vastly more time and resource consuming than the proposed strategy. Mean and total runtimes for both Liu *et al.* [33, 34] followed by a Wiener Filter [35] and our method for all 160 ROIs selected from the DDSM are shown in Table II.

The total runtime for the execution of BPPM in all images using 7 different values for the parameter K is 1089 seconds, or 6 times faster than the 6564 seconds taken by Liu *et al.* [33, 34]. Therefore, if processing time and/or memory consumption are limited – as is the case of on-chip implementations – one could easily embed our algorithm in a microchip and perform the same z-score analysis with only one small context (e.g., $K = 2$) and one large context (e.g., $K = 10$), still obtaining consistent results. Another advantage of our method is that, once the number of noisy BPs is known using z-score analysis for a given sensor, a simple pixel-wise quantization with order $\Theta(n)$ can be performed in order to achieve noise removal of a new image from the same sensor. In comparison, each new

TABLE II
RUNTIMES IN SECONDS OF LIU *et al*. [33, 34] + WIENER FILTER [35] AND BPPM [19] + BP REMOVAL.

		160 ROIs	Mean Runtime per Image
Liu *et al*. [33, 34] + Wiener Filter [35]		6564	41.03
BPPM [19] + BP Removal	k_1	155	0.97
	k_2	151	0.94
	k_3	152	0.95
	k_4	152	0.95
	k_5	152	0.95
	k_{10}	155	0.97
	k_{15}	172	1.08
	Total	1089	6.80

image processed by [33, 34] needs to have its noise estimated and processed by a Wiener Filter [35].

The second method used to validate the removal of noisy BPs was an accuracy analysis from the Machine Learning algorithms presented in Section III, as shown in Fig. 5. It is noteworthy that most of those algorithms are non-deterministic and therefore required replication steps to achieve good estimates for their accuracies. Each algorithm was executed 5 times and an average of their accuracies was computed. For all classifiers in Fig. 5, the x-axis corresponds to the amount of BPs maintained in the radiometric resolution of the images and the y-axis presents the correspondent algorithm accuracy in the task of assessing the malignancy of breast nodules. The largest accuracies were achieved while using the SVM and DT algorithms. SVM reached 88.73% when 8 BPs were maintained and 89.31% when 6 BPs were preserved, while DT achieved 88.92% when 7 BPs were maintained. The accuracies achieved in our tests do not comprise the state-of-the-art of breast nodule classification, however they serve as evidence for the suboptimality of texture features in noisy channels.

As can be seen in Fig. 5, the accuracy of most algorithms is larger when 4 or 5 LSBPs were removed from the DDSM ROIs prior to texture extraction in our analysis. If more than 6 of the LSBPs are removed from the images, accuracies tend to decrease. This effect is expected, as MSBPs tend to store useful information, that is, data less affected by noise. Fig. 6 shows the differences in accuracy when using the whole images (12 b/p) and when using only 9, 8, 7 or 6 b/p. Visual inspection reveals that when after the removal of 4 and 5 BPs – leaving the images with 8 and 7 b/p respectively – almost all classifiers achieved better accuracies, with the exception of ET with 7 b/p. Although SVM and NB peaked when using 6 b/p, classifiers DT, RF and ET showed losses in accuracy, indicating that the use of 7 or 8 b/p for these images results in more consistent gains.

This paper's results imply that texture-based classification methods for DDSM images should perform a quantization step to remove between 3 and 5 BPs previously to feature extraction. Images acquired using newer imaging techniques for mammography – as FFDMs and Computerized Radiography (CR) mammograms – should present less BPs affected by noise and, therefore, less BPs should be statistically indistinguishable from randomly generated noise.

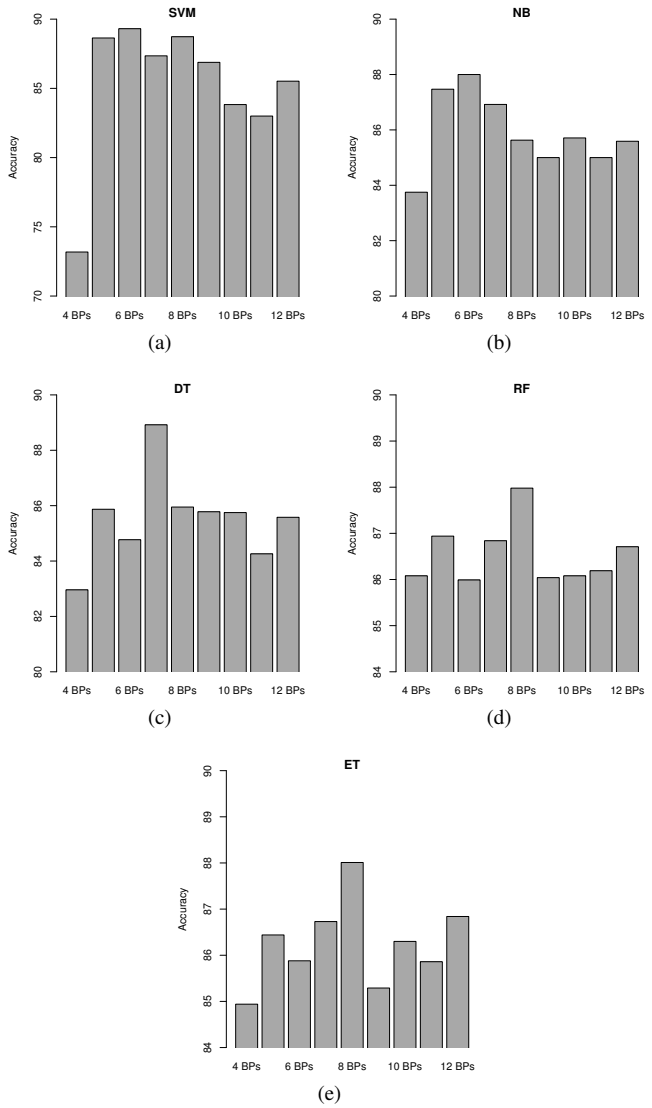

Fig. 5. Accuracies of 5 Machine Learning algorithms using different numbers of BPs. (a) SVM. (b) NB. (c) DT. (d) RF. (e) ET.

V. DISCUSSION AND CONCLUSION

This paper proposes an investigation on the influence of noise in malignancy classification of mammographic findings. Even though breast cancer has good prognosis is most cases, it is crucial to detect this disease in the early stages of its development. Computer applications like mammography CAD systems assist radiologists in detection and diagnosis of breast cancer can be used as a second opinion. According to Moreira *et al*. [12] DDSM is the database most used in literature. However, mammographic images contain large amounts of noise and many works [3, 13, 14, 15] do not perform proper noise removal preprocessing.

In order to analyse the amount of noise from biomedical images, we developed a new method based on Information Theory and Data Compression for BP random noise detection

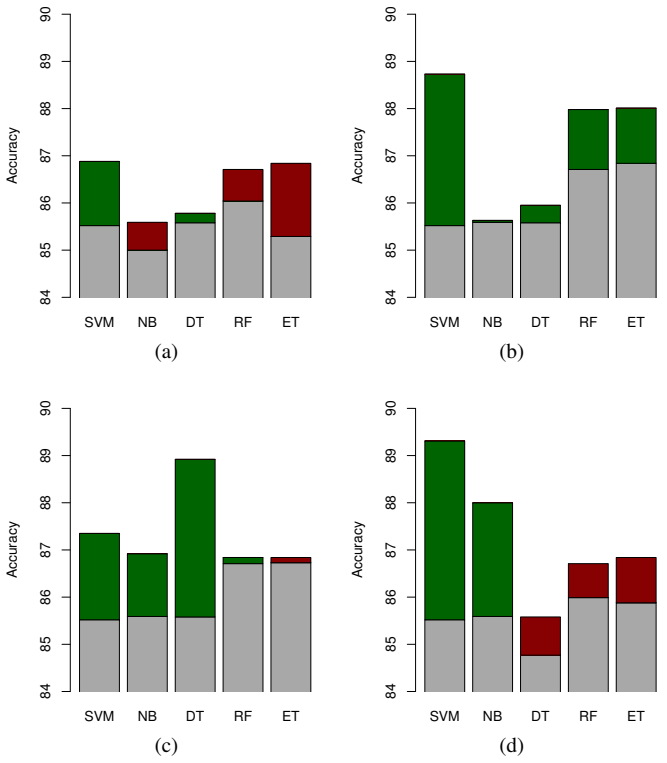

Fig. 6. Differences of accuracies for 5 Machine Learning algorithms. Gains in accuracy when removing BPs are shown in green while losses are represented by red stacker bars. (a) 12 b/p vs. 9 b/p. (b) 12 b/p vs. 8 b/p. (c) 12 b/p vs. 7 b/p. (c) 12 b/p vs. 6 b/p.

and removal. The method was validated using the noise assessment method NoiseLevel [33, 34] in DDSM images. A second validation to our results was performed using Machine Learning algorithms on quantized images and analysing their accuracies. Fig. 5 shows gains in accuracy of classifiers as the radiometric resolution is reduced by removal of LSBPs, confirming the negative influence of noise in DDSM images with 12 BPs. However, tests wherein more than 5 BPs were removed presented drastic decreases in accuracy for most algorithms, implying that important texture data is lost with the removal of these BPs.

One advantage of our method, when compared to other state-of-the-art noise removal strategies like Vieira *et al.* [16] is that BP removal – which can also be understood as a scalar quantization – do not insert artificial information on pixels, but instead only removes noisy, unnecessary and potentially hampering data. In addition, many strategies do their processing steps in the frequency domain, while all our algorithm is executed in space domain, lowering its computational requirements. Therefore, if our noise removal method needs to be executed in an embedded system, as a camera or mammographic scanner, all its stages are friendly to hardware implementation. Discarding BPs highly infected by random noise also has the indirect beneficial effect of compressing the image without losing important information.

The methodology proposed in this paper showed evidences that corroborate the necessity of reduction in radiometric resolution in noisy digitized mammograms. Future works include validation of the proposed methodology in other datasets, mammographic imaging techniques and scanners; and in other types of biomedical data, as images obtained from Mammary Tomosynthesis, Magnetic Resonance Imaging, Computerized Tomography and temporal signals, as Electrocardiograms and Electroencephalograms.

Acknowledgment

The authors would like to thank CAPES for the financial support provided during the making of this article.

References

[1] A. Jemal, R. Siegel, E. Ward, T. Murray, J. Xu, and M. J. Thun, "Cancer statistics, 2007," *CA: a cancer journal for clinicians*, vol. 57, no. 1, pp. 43–66, 2007.

[2] T. Tan, B. Platel, R. Mus, L. Tabar, R. M. Mann, and N. Karssemeijer, "Computer-aided detection of cancer in automated 3-d breast ultrasound," *Medical Imaging, IEEE Transactions on*, vol. 32, no. 9, pp. 1698–1706, 2013.

[3] X. Li, S. Williams, and M. J. Bottema, "Texture and region dependent breast cancer risk assessment from screening mammograms," *Pattern Recognition Letters*, vol. 36, pp. 117–124, 2014.

[4] S. Yu and L. Guan, "A cad system for the automatic detection of clustered microcalcifications in digitized mammogram films," *Medical Imaging, IEEE Transactions on*, vol. 19, no. 2, pp. 115–126, 2000.

[5] A. Jalalian, S. B. Mashohor, H. R. Mahmud, M. I. B. Saripan, A. R. B. Ramli, and B. Karasfi, "Computer-aided detection/diagnosis of breast cancer in mammography and ultrasound: a review," *Clinical imaging*, vol. 37, no. 3, pp. 420–426, 2013.

[6] M. J. G. Calas, B. Gutfilen, and W. C. d. A. Pereira, "Cad and mammography: why use this tool?" *Radiologia Brasileira*, vol. 45, no. 1, pp. 46–52, 2012.

[7] Y. Lan, H. Ren, Y. Zhang, and H. Yu, "Similarity in mammography cad using cbir approach: A validation study," in *Intelligent Human-Machine Systems and Cybernetics (IHMSC), 2012 4th International Conference on*, vol. 2. IEEE, 2012, pp. 373–376.

[8] S. Timp, C. Varela, and N. Karssemeijer, "Computer-aided diagnosis with temporal analysis to improve radiologists' interpretation of mammographic mass lesions," *Information Technology in Biomedicine, IEEE Transactions on*, vol. 14, no. 3, pp. 803–808, 2010.

[9] L. Tabar, C. Fagerberg, A. Gad, L. Baldetorp, L. Holmberg, O. Grontoft *et al.*, "Randomised trial from the breast cancer screening working group of the swedish national board of health and welfare. reduction in mortality from breast cancer after mass screening with mammography," *Lancet*, vol. 1, pp. 829–32, 1985.

[10] D. A. Berry, K. A. Cronin, S. K. Plevritis, D. G. Fryback, L. Clarke, M. Zelen, J. S. Mandelblatt, A. Y. Yakovlev,

J. D. F. Habbema, and E. J. Feuer, "Effect of screening and adjuvant therapy on mortality from breast cancer," *New England Journal of Medicine*, vol. 353, no. 17, pp. 1784–1792, 2005.

[11] Z. Wang, G. Yu, Y. Kang, Y. Zhao, and Q. Qu, "Breast tumor detection in digital mammography based on extreme learning machine," *Neurocomputing*, vol. 128, no. 0, pp. 175–184, 2014.

[12] 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.

[13] N. Dhungel, G. Carneiro, and A. P. Bradley, "Automated mass detection in mammograms using cascaded deep learning and random forests," in *Digital Image Computing: Techniques and Applications (DICTA), 2015 International Conference on*. IEEE, 2015, pp. 1–8.

[14] Y. Chen, Y. Lan, and H. Ren, "A feature selection method base on ga for cbir mammography cad," *Intelligent Human-Machine Systems and Cybernetics (IHMSC)*, vol. 2, pp. 175–178, 2012.

[15] A. u. Rehman, N. Chouhan, and A. Khan, "Diverse and discriminative features based breast cancer detection using digital mammography," in *Frontiers of Information Technology (FIT), 2015 International Conference on*. IEEE, 2015, pp. 234–239.

[16] M. A. Vieira, P. R. Bakic, A. D. Maidment, H. Schiabel, and N. D. Mascarenhas, "Filtering of poisson noise in digital mammography using local statistics and adaptive wiener filter," in *Breast Imaging*. Springer, 2012, pp. 268–275.

[17] C. E. Shannon, "A mathematical theory of communication," *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 5, no. 1, pp. 3–55, 2001.

[18] J. G. Cleary and I. H. Witten, "Data compression using adaptive coding and partial string matching," *IEEE Transactions on Communications*, vol. 32, no. 4, pp. 396–402, Apr. 1984.

[19] J. R. Marques, G. M. Pires, L. V. Batista, and J. Poel, "Compressão de imagens mamográficas utilizando segmentação e o algoritmo ppm," in *X Congresso Brasileiro de Informática em Saúde*, 2006, pp. 22–27.

[20] J. J. Brasileiro and A. C. Cavalcanti, "On-chip integration of a binary ppm algorithm for lossless compression of ecg: Design space exploration for low hardware complexity," *II Workshop on Circuits and System Design*, 2012.

[21] M. Heath, K. Bowyer, D. Kopans, R. Moore, and W. P. Kegelmeyer, "The digital database for screening mammography," in *Proceedings of the 5th international workshop on digital mammography*. Citeseer, 2000, pp. 212–218.

[22] V. Gaike, N. Akhter, K. Kale, and P. Deshmukh, "Application of higher order glcm features on mammograms," in *Electrical, Computer and Communication Technologies (ICECCT), 2015 IEEE International Conference on*. IEEE, 2015, pp. 1–3.

[23] D. C. Moura and M. A. G. López, "An evaluation of image descriptors combined with clinical data for breast cancer diagnosis," *International journal of computer assisted radiology and surgery*, vol. 8, no. 4, pp. 561–574, 2013.

[24] D. Wang, L. Shi, and P. A. Heng, "Automatic detection of breast cancers in mammograms using structured support vector machines," *Neurocomputing*, vol. 72, no. 13, pp. 3296–3302, 2009.

[25] R. M. Haralick, K. Shanmugam, and I. H. Dinstein, "Textural features for image classification," *Systems, Man and Cybernetics, IEEE Transactions on*, no. 6, pp. 610–621, 1973.

[26] U. Acharya, S. Vinitha Sree, L. Saba, F. Molinari, S. Guerriero, and J. Suri, "Ovarian tumor characterization and classification: A class of gynescan systems," in *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE*. IEEE, 2012, pp. 4446–4449.

[27] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[28] A. Harisinghaney, A. Dixit, S. Gupta, and A. Arora, "Text and image based spam email classification using knn, naive bayes and reverse dbscan algorithm," in *Optimization, Reliabilty, and Information Technology (ICROIT), 2014 International Conference on*. IEEE, 2014, pp. 153–155.

[29] N. Azizi, N. Zemmal, M. Sellami, and N. Farah, "A new hybrid method combining genetic algorithm and support vector machine classifier: Application to cad system for mammogram images," in *Multimedia Computing and Systems (ICMCS), 2014 International Conference on*. IEEE, 2014, pp. 415–420.

[30] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and regression trees*. CRC press, 1984.

[31] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine learning*, vol. 63, no. 1, pp. 3–42, 2006.

[32] "DDSM Website," http://marathon.csee.usf.edu/Mammography/Database.html, accessed: 2016-07-20.

[33] X. Liu, M. Tanaka, and M. Okutomi, "Noise level estimation using weak textured patches of a single noisy image," in *Image Processing (ICIP), 2012 19th IEEE International Conference on*. IEEE, 2012, pp. 665–668.

[34] ——, "Single-image noise level estimation for blind denoising," *Image Processing, IEEE Transactions on*, vol. 22, no. 12, pp. 5226–5237, 2013.

[35] M. Kazubek, "Wavelet domain image denoising by thresholding and wiener filtering," *IEEE Signal Processing Letters*, vol. 10, no. 11, pp. 324–326, 2003.