Investigation of Classification Algorithms for a Prototype Microwave Breast Cancer Monitor

Adam Santorelli, Yunpeng Li, Emily Porter, Milica Popović, Mark Coates
Department of Electrical Engineering, McGill University, Montreal, Canada
{adam.santorelli, yunpeng.li, emily.porter}@mail.mcgill.ca, {milica.popovich, mark.coates}@mcgill.ca

***Abstract*—In this paper we investigate the use of differential signals to monitor changes within the breast. Specifically, we focus on the use of machine learning classification algorithms to determine whether any malignant tissues are developing. Experimental data is obtained from a 16-element antenna array that transmits a 2 – 4 GHz broadband pulse. We implement both the Linear Discriminant Analysis and Support Vector Machine (SVM) detection algorithms to analyze the experimentally obtained data.**

***Index Terms*— cancer detection; classification algorithms; microwave sensors.**

I. INTRODUCTION

The early detection of breast cancer is pivotal to ensure successful treatment [1]. Microwave systems offer the possibility of a complementary modality to the current standard of X-ray mammography for breast cancer screening [1]. These systems offer the advantage of pain-free non-ionizing scans. The ability to create affordable microwave systems would allow for scans to be easily available to an increased number of women; additionally, the non-ionizing nature of the scan enables the possibility of more frequent exams. These factors make microwave systems a strong candidate for breast cancer monitoring.

Recent literature has shown that machine learning algorithms can be applied to data obtained from microwave systems in order to classify and detect the presence of a tumour embedded within the breast [2-4]. In [4], it was shown that applying the support vector machine (SVM) classifier to numerical differential signals, signals obtained from successive scans of the breast over time, improved classification accuracy in a heterogeneous breast scenario. These results suggest that frequent scans with a microwave system can be used to detect changes within the breast tissue, and classify these changes as benign or cancerous. However, these papers have only focused on numerically obtained data. Most recently, we have demonstrated that these classification algorithms, both SVM and linear discriminant analysis (LDA), can be applied to data obtained from realistic breast phantoms using an experimental time-domain system [5].

In this paper, we apply a machine learning classification algorithm to experimentally obtained data to monitor changes within the breast. We use a 16-element antenna array to perform measurements in the time-domain on dielectrically-realistic breast phantoms. Unlike in [5], we now investigate the use of differential signals from successive scans of the breast phantoms over time. In an experimental system, variation of both the direct pulse signal (signals travelling directly between antennas) and the background environment signal degrades the classifier performance as these variations are not associated with changes within the breast structure. This results in a direct pulse residual after subtraction. We minimize this residual by following the calibration procedure proposed in [6]. We apply a machine learning algorithm to differential signals obtained from our time-domain system and evaluate the ability of the system to monitor changes within the breast structure.

II. METHODOLOGY

A. System Overview and Data Collection

A complete description of our time-domain system can be found in [7]. The system is composed of a 16-element antenna array, specifically designed as a microwave sensor for breast cancer detection, and held in place by a hollow hemi-spherical bowl-shaped radome into which the breast phantom under test is placed. A pulse with spectral content in the 2 – 4 GHz range is fed into a 16x2 switching matrix that chooses the specific transmitting and receiving antenna pairs, such that, for each scan, a total of 240 bistatic signals are recorded by an oscilloscope with an equivalent time-sampling rate of 80GSa/s. We record 4096 samples; however, during pre-processing, data is windowed, based on the longest possible path for a wave to travel, to only include 320 samples (4ns). A photograph of the experimental system is shown in Figure 1.

We fabricated tissue phantoms, following the procedure proposed in [8], to closely mimic the dielectric properties of actual tissues based on the measurements found in [9]. Specifically, in this study we focused on homogeneous phantoms that mimic adipose tissue. However, despite following the same fabrication procedure, significant variation in the dielectric properties, from phantom to phantom and within a singular phantom, can be seen [10]. This variation increases the complexity of the problem and approaches the real world problem where no two breasts have the exact same composition.

Multiple breast scans were taken from the same phantom over a period of approximately two weeks. As was shown in [5] and [8] the dielectric properties of the breast phantom vary over time. In [5], measurements over a period of 7 days demonstrated that the relative permittivity, ε_r, at 3 GHz of the healthy breast phantom varied between 5.5 and 19.

This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC), le Fonds québécois de la recherche sur la nature et les technologies (FQRNT) and Partenariat de Recherche Orientée en Microélectronique, Photonique et Télécommunications (PROMPT).

978-88-907018-4-9/14/$31.00 ©2014 IEEE

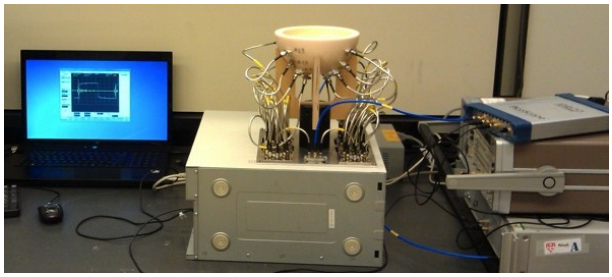

Figure 1: A photograph of the time-domain microwave system.

For "healthy" breast phantoms, the radome is completely filled with the adipose tissue. To represent a "tumourous" phantom, a spherical 1 cm radius tumour (a tumour of this size represents the largest possible tumour that can be categorized as "Stage 1" breast cancer) is carved from fabricated tumour phantoms. The tumour is then embedded in the healthy breast phantom in one of two possible locations; Position A at a depth of 2 cm and located halfway between the center and radome wall, and Position B at a depth of 3 cm and located at a 180° rotation from Position A. An illustration of the two possible tumour locations is shown in Figure 2.

The differential signals are obtained by comparing successive breast scans to the initial breast scan. Thus, we are able to monitor changes within the breast structure that are either due to slight variations of the healthy breast tissue (variation of the phantom properties over time) or the presence of a tumour (embedding the tumour phantom into the healthy breast phantom). Tumourous differential signal sets are defined as the difference between breast scans performed with a tumourous and a healthy breast phantom, whereas a healthy differential data set is the difference between scans of the same healthy breast phantom.

B. Detection Algorithm

In this section we describe the steps followed to successfully implement the proposed method. Similar to [4], [5], we apply both LDA and SVM classifiers, individually, on selected data features that are extracted using Principle Component Analysis (PCA).

1) Noise Mitigation

Inherent in the experimental data are sources of vertical and horizontal noise. The primary source of vertical noise is due to white noise from the limitations of the measurement equipment (oscilloscope). Horizontal noise arises due to jitter in the clock and oscilloscope trigger signal. To mitigate these sources of noise we use a correlation alignment procedure. More detail is documented in [10]. Additionally, we follow the method proposed in [6] to minimize the direct pulse residual.

2) Differential Signals

A total of 150 breast scans from 10 breast phantoms were used to obtain the training data set. Let $\boldsymbol{X_m}$ be the complete data set recorded from a specific breast phantom m, where $m = 1, \ldots, 10$. We can define the k^{th} differential signal data set, $\Delta\boldsymbol{X}_{mk}$, as the difference between the initial breast scan $\boldsymbol{X}_{m1}$ and some future breast scan, $\boldsymbol{X}_{mj}$, of the same phantom, such that:

$$\Delta\boldsymbol{X}_{mk} = \boldsymbol{X}_{mj} - \boldsymbol{X}_{m1}$$
$$for\ j = 2, 3, \ldots, n\ and\ k = 1, 2, \ldots, n-1 \quad (1),$$

where n is the total number of breast scans performed on each specific phantom m. This process is repeated for each of the 10 breast phantoms; thus, we obtain 140 differential data sets, of 240 signals each, to train the classifier. The training data is evenly divided (70 data sets each) between healthy differential data sets and tumourous differential signals. Of the 70 data sets representing a tumourous differential, half are obtained from scans when a tumour is embedded in Position A and the other half from Position B.

The aim of this study is to mimic a real-life scenario in which we make a decision on a new patient based on information from other patients of similar breast size. In this case the training data represents scans from individuals whom we know to have either healthy or unhealthy breasts; the testing data represents a scan of a new patient. Thus, to ensure there is no overlap between the training and testing data, the test data set is obtained from two breast phantoms not included in the initial 10 breast phantoms used to train the classifier. The first scan of each new breast phantom is used as the reference ($\boldsymbol{X_1}$), and each subsequent scan is used to obtain the differential data set, $\boldsymbol{\Delta X_k}$, for each phantom. The test data set is made up of 32 scans, equally distributed between scans when no tumour has developed (healthy) and when a tumour is embedded at one of the two positions; i.e. 16 scans where no tumour is inserted into the phantom, 8 cases where the tumour is inserted into the phantom at Position A, and 8 cases when the tumour is inserted into Position B.

An example of the three types of differential signals for a specific antenna pair (Antenna 1 transmits, Antenna 3 receives) is shown in Figure 3. These differential signals represent the difference between two successive scans of a breast phantom when (i) there are no changes to the tissue, shown in blue, (ii) a tumour is embedded in the phantom at Position A, shown in red, and (iii) a tumour is embedded at Position B, shown in green. From Figure 2 we observe that, starting at about 50 samples, there is a region of interest (highlighted in the plot with the dashed box) where three signal types are very different from each. Thus, prior to any further analysis, the data set is windowed to only retain information after the first 50 samples.

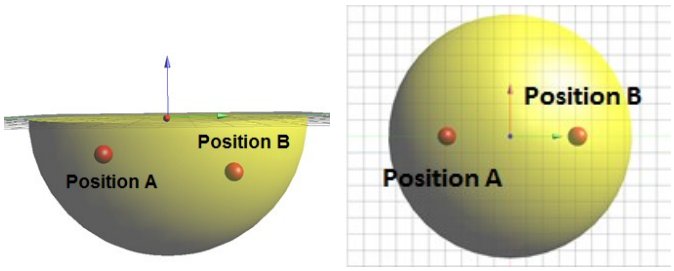

Figure 2. An illustration showing the breast phantom (yellow), and the two possible tumour locations, denoted by the red spherical tumours. A side (left) and top (right) view is shown.

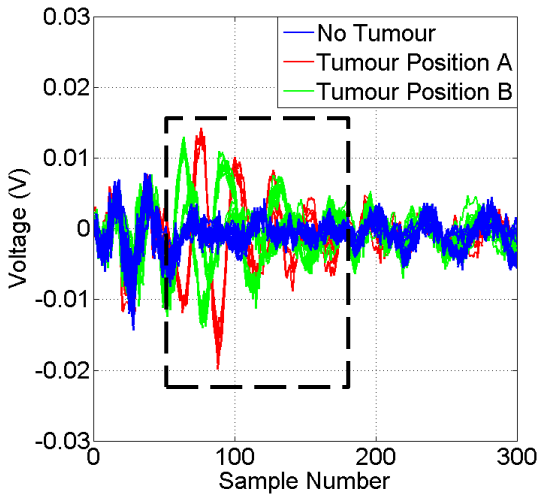

Figure 3: A comparison of the three types of differential signals obtained for a specific antenna pair (Antenna 1 transmits, Antenna 3 receives). We highlight a region of interest where the three differential signals are most varied.

3) *Feature Extraction and Classifier Optimization*

The dimensionality of the data can be reduced by using feature extraction and retaining only a select number of features. This procedure allows us to retain the most important features of the data set while also reducing the influence of less relevant information such as noise. We use Principle Component Analysis (PCA) to extract these features from the data. The classification algorithms are then applied to these extracted features. The choice of using PCA for feature extraction is based on the results in [12], where it was shown to be the best feature extraction method when paired with SVM and on average. The number of principle components to retain is chosen based on the results from cross-validation, described in detail later in this section. A difference from our past work in [5] is that we no longer standardize the data before feature extraction due to the variation of the signal strength throughout the data; in fact, there are sections of the data where the signals are extremely weak, on the order of noise.

A support vector machine (SVM) is a machine learning algorithm that is best used when the data has exactly two classes. The training data is mapped to a higher dimensional space and used to create support vectors such that a discriminant hyperplane that separates the two classes can be created. The SVM classifier maximizes the distance (margin) between this hyperplane and the training data set. The LDA, an easier and more simply implemented classification algorithm, is used as a benchmark to compare with the performance of SVM [3] – [5].

The performance of the SVM classifier is optimized by choosing specific operating parameters; in particular, we must determine the optimal box constraint, C, and the radial-basis function sigma value, σ, as well as the number of principle components, n, [5]. As described in [12], we use a coarse-the-fine grid search for these operating parameters. Additionally, we perform 10-fold cross-validation to determine these optimal parameter values based on the correct classification rates from cross-validation. These parameters are then applied to the final testing scenario. From our 10-fold cross validation we find that the optimal number of principle components is 25 and the box constraint and sigma values are 10 and 4, respectively.

4) *Data Fusion*

The above mentioned analysis is done on a signal-to-signal basis. All the data recorded from each specific antenna pair, over all the breast scans performed, are grouped together. Feature extraction and classification is then applied to this grouped data. Thus, for each breast scan we must perform PCA and then classification 240 times. The resulting output of the classification step is then 240 'decisions'; for each antenna pair the classifier gives a decision of 'healthy' or 'tumour'.

In [5] we proposed, based on observations in previous work, that certain antenna pairs may be more sensitive to changes within the breast tissue. Specifically, we investigated the idea that the four antennas closest to each other (whose signals travel the least distance within the breast prior to communicating with each other) would have the best performance. We observed that only using the data from these 48 antenna pairs (4 antennas in each of the 4 quadrants) improved the sensitivity to presence of the tumour but increased the false-positive rates [5].

In the here reported work, we aim to take into account the recordings of these antenna pairs simultaneously; that is to say we will group the data from these antenna pairs *prior* to feature extraction (data from each antenna pair within a group are concatenated). Thus, we make use of the information of the group of antenna as a whole. Additionally, we do not discard the remaining antenna pairs.

Each breast scan is composed of 240 recorded signals, representing the 240 antenna pairs. We group these 240 signals into 20 groups of 12 antenna pairs. Four of these groups, 48 antenna pairs, represent the four quadrants of the antenna array. The remaining 192 antenna pairs are randomly distributed into 16 groups of 12. Thus, we need only perform feature extraction and classification 20 times per breast scan.

5) *Data Averaging and Thresholding*

The output of the classification algorithm is a numerical value, a 0 or 1 depending on the decision made, healthy or tumour respectively. The total number of decisions per breast scan is equal to the number of antenna pairs, or groups, used in the analysis. We propose that it is possible to use information from each decision made to determine the type of phantom, or breast, being scanned. We can average the result from each antenna pair, or group of antenna pairs if we are using data fusion, and produce a single numerical value representing each specific breast scan. This numerical value is hereafter referred to as NV. Depending on the NV value, we can make a decision on whether the signals collected for that specific scan represent changes within the breast that are benign or malignant.

In this paper, we will discuss the possibility of using a threshold value, t_h, such that if $NV > t_h$, the breast scan contains a tumour, and if $NV < t_h$, the scan represents a healthy breast

phantom. The aim in this paper is to simply present a range of threshold values, such that adopting any threshold value within that range guarantees a 100% correct decision on the nature (healthy or unhealthy) of the phantom being scanned. Additionally, we will show that using data fusion will increase the range that leads to 100% correct decision making.

Future work will focus on how to choose this threshold value and whether the choice of this t_h value should favor false-positive or false-negative rates, or some medium in between.

III. RESULTS

The LDA classifier can be implemented without any optimization, while the SVM classifier must be trained with the optimal parameter values (n, C, and σ) found from the optimization procedure. The trained classifiers are tested with the testing data set obtained from the two newly fabricated breast phantoms.

In Table 1 we compare the optimized-SVM classifier by presenting the overall correctly classified percentage, the false-positive, and the false-negative rates; additionally, we present the range of threshold values for which we can guarantee a 100% correct decision will be made based on the averaged NV values. Results without using differential signals, as presented in [5], are shown in parenthesis. When using the differential signal analysis the overall correctly classified percentage is decreased by 8.6% for the LDA classifier, while it is increased by 3.89% for the SVM classifier.

The results from the data fusion technique outlined in Section II.B.4 are shown in Table 2. We compare these results to a naïve form of grouping in order to assess performance of the data fusion technique. Data fusion groups data *prior* to feature extraction and classification, whereas the naïve grouping is performed *after* analysis. The naïve grouping analysis is performed by applying the same grouping as implemented in the data fusion technique, but we now use a majority-vote on the output of the classifier to obtain a group decision. The naïve grouping results are shown in parenthesis in Table 2.

In comparison to the results presented in Table 1, using data fusion provides an increase in performance for both classifiers. The overall detection rate is greatly improved for both classifiers, and the difference in performance between the two classifiers is greatly reduced. Both the false-positive and false-negative rates for both classification algorithms are greatly decreased. Additionally, it is clear that using the data-fusion technique provides an increased improvement in comparison with a naïve grouping analysis. We also note that the naïve grouping widens the gap between the false-positive and false-negative rates, in comparison with the no-grouping and data-fusion techniques.

We observe that the threshold range that leads to 100% correct decision making is significantly increased, thus reducing the burden for choosing the appropriate threshold value for decision making. Clearly, the data fusion technique is very useful in improving the performance of the classification algorithms. Furthermore, we conclude that using the data-fusion technique in conjunction with a simple and easily implemented classification algorithm such as LDA, which requires no optimization, can be used to obtain correct classification results of almost 90%, with a false-negative rate of almost 5%. Additionally, we note that the range of threshold values for the LDA classifier with data fusion spans from 0.222 to 0.944, suggesting that decision making based on information from all the antenna groups can be quite successful.

TABLE I. A COMPARISON OF THE CLASSIFICATION RESULTS FOR LDA AND SVM WITH AND WITHOUT[5] THE USE OF DIFFERENTIAL SIGNALS. RESULTS FROM [5] SHOWN IN PARENTHESIS.

	Correctly Classified (%)	**False-Positive (%)**	**False-Negative (%)**	**Threshold Value Range**
LDA	61.70 (70.30)	66.39 (30.00)	10.27 (29.55)	[0.771, 0.862]
SVM	77.53 (73.64)	8.33 (32.52)	37.41 (23.29)	[0.091, 0.605]

TABLE II. CLASSIFICATION RESULTS FOR LDA AND SVM USING DATA FUSION TECHNIQUE FOR SPECIFICALLY GROUPING 12 ANTENNA PAIRS COMPARED WITH NAÏVE DATA GROUPING IN PARENTHESIS

	Correctly Classified (%)	**False-Positive (%)**	**False-Negative (%)**	**Threshold Value Range**
LDA	89.58 (58.51)	15.28 (78.82)	5.56 (4.17)	[0.222, 0.944]
SVM	91.15 (80.73)	5.56 (2.43)	12.15 (36.11)	[0.056, 0.833]

IV. CONCLUSION

We demonstrate that both an LDA and SVM classifier can be used to detect changes within the breast structure using differential signals from successive breast scans obtained from an experimental time-domain microwave system; additionally, we note that using differential signals improves detection performance for the SVM classifier. We demonstrate that a data fusion technique, which uses information from groups of antenna, can greatly improve the performance of both classifiers by reducing both false-positive and false-negative rates. We propose the idea that a threshold value can be used to make a decision on the type of breast, or phantom, being scanned by taking into account information from all classification decisions (from each antenna pair, or group of antenna pairs).

Future work involves investigating whether such a system can detect healthy changes in glandular content and differentiate this from the development of breast tumours. We aim to investigate if grouping more antenna pairs together can improve classification results further, and how to go about choosing a threshold value such that we can make a decision on the nature of the recorded signals.

REFERENCES

[1] Nikolova, N., "Microwave Imaging for Breast Cancer," *IEEE Microwave Magazine*, pp. 78 – 94, Dec. 2011.

[2] S. Davis,, et al. "Breast tumor characterization based on ultrawideband microwave backscatter." *Biomedical Engineering, IEEE Transactions on* 55.1 (2008): 237-246.

[3] R. C. Conceição, M. O'Halloran, M. Glavin, and E. Jones, "Support Vector Machines for the Classification of Early-stage Breast Cancer Based on Radar Target Signatures," *Prog. Electromagn. Res. B*, vol. 23, pp. 311–327, 2010.

[4] D. Byrne, M. O'Halloran, M. Glavin, and E. Jones, "Breast cancer detection based on differential ultrawideband microwave radar," *Prog. Electromagn. Res. M*, Vol. 20, 231-242, 2011.

[5] A. Santorelli, E. Porter, E. Kirshin, Y. J. Liu, and M. Popovic, "Investigation of classifiers for tumor detection with an experimental time-domain breast screening system," *Progress In Electromagnetics Research*, Vol. 144, 45-57, 2014

[6] A. Santorelli, E. Kirshin, E. Porter, M. Popovic, J. Schwartz, "Improved calibration for an experimental time-domain microwave imaging system," *Antennas and Propagation (EuCAP), 2013 7th European Conference on* , vol., no., pp.801,805, 8-12 April 2013.

[7] E. Porter, E. Kirshin, A. Santorelli, M. Coates, and M. Popović, "Time-domain multistatic radar system for microwave breast screening," *IEEE Antennas Wireless Propag. Lett.*, vol. 12, pp. 229-232, 2013.

[8] E. Porter, J. Fakhoury, R. Oprisor, M. Coates, and M. Popović, "Improved Tissue Phantoms for Experimental Validation of Microwave Breast Cancer Detection", *Antennas and Propagation (EUCAP), Proceedings of the 4th European Conference on*, Barcelona, Spain, 12-16 April 2010.

[9] M. Lazebnik, et al., "A large-scale study of the ultrawideband microwave dielectric properties of normal, benign and malignant breast tissues obtained from cancer surgeries," *Phys. Med. Biol*., Vol. 52, pp. 6093-6115, 2007.

[10] E. Porter, E. Kirshin, A. Santorelli, and M. Popovic, "Microwave breast screening in the time-domain: identification and compensation of measurement-induced uncertainties," *Progress In Electromagnetics Research B*, Vol. 55, 115-130, 2013.

[11] C.-W Hsu, C.-C. Chang, and C.-J. Lin, "A practical guide to support vector classification," Department of Computer Science, National Taiwan University, Tech. Rep., 2003. [Online]. Available: http://www.csie.ntu.edu.tw/ cjlin/papers.html.

[12] R.C., Conceicao, et al. "Evaluation of features and classifiers for classification of early-stage breast cancer." *Journal of Electromagnetic Waves and Applications* 25.1 (2011): 1-14.