COST-SENSITIVE ENSEMBLE CLASSIFIERS FOR MICROWAVE BREAST CANCER DETECTION

Yunpeng Li, Adam Santorelli, Olivier Laforest and Mark Coates

Dept. of Electrical and Computer Engineering, McGill University, Montréal, Québec, Canada
{yunpeng.li, adam.santorelli, olivier.laforest}@mail.mcgill.ca, mark.coates@mcgill.ca

ABSTRACT

Microwave breast cancer detection involves analysing the scattered waveforms of microwave signals that are propagated into the breast. We have developed a microwave-radar time-domain system and performed clinical trials using a prototype. This paper presents a classification architecture based on cost-sensitive support vector machines that is designed to process the signals measured by the 16-element multi-static antenna array. We examine the performance of the classifier by applying it to measurements performed on tissue-mimicking breast phantoms.

***Index Terms*—** Microwave breast cancer detection, ensemble classifier, support vector machine, Neyman-Pearson classification

1. INTRODUCTION

Early detection of breast cancer significantly improves the chance of successful treatment of the disease. Currently, the most prevalent and effective breast cancer screening method is x-ray mammography [1]. It has several drawbacks, including the use of ionizing radiation, uncomfortable breast compression, and high miss probability. Ultrasound and magnetic resonance imaging can provide complementary information but also have disadvantages [2]. Microwave breast cancer detection has the potential to act as a valuable complementary modality; it is based on the reported inherent contrast between the dielectric properties of malignant and healthy breast tissues [3]. Scattering arises at regions of significant contrast in dielectric properties. Tomographic imaging methods aim to reconstruct a dielectric profile of the breast tissue [4]; radar methods try to map regions of dielectric scattering, from which tissue types can be inferred.

Numerous signal processing approaches have been proposed for analysing the signals obtained by microwave systems [2], but most have not been assessed experimentally, so many ignore critical practical challenges. Imaging techniques are the most common; these focus on creating an image that can be later assessed by a clinical expert. Imaging has been performed using delay-and-sum and beamforming methods [5–10] and hypothesis test-based methods [11]. Most of these techniques require accurate models of the wave propagation delays (and in some cases even the scattered signals), implying knowledge of tissue characteristics and skin thickness, which is unavailable in practice.

Recently, some research has explored the application of supervised learning algorithms to signals recorded by microwave breast cancer systems. In [12], Byrne et al. applied a support vector machine (SVM), using features extracted by principal component analysis (PCA). Santorelli et al. extend this idea in [13] by utilizing a data fusion strategy to boost the classifier performance. In both cases, classifiers are applied to individual signals and there is no attempt to merge the decisions to provide a detection output for the entire breast. There is no systematic procedure for appropriately choosing thresholds to control the false alarm or false discovery rate.

The main aim in microwave breast cancer detection is to provide an easy-to-use, inexpensive, and safe early warning system. A positive response from the system indicates that the patient should undergo more comprehensive conventional tests using other modalities. It is important to achieve the best detection performance while controlling the rate at which tumour-free breasts are mistakenly classified as requiring further tests. In this paper, we develop classification architectures that jointly process all of the signals measured in a breast scan and provide a single decision as to whether further tests are necessary. We use a cost sensitive SVM technique, the 2ν-SVM [14], to conduct microwave breast cancer detection in the Neyman-Pearson (NP) context.

The paper is organized as follows. Section 2 formalizes the problem and Section 3 presents novel cost-sensitive ensemble classifiers. Their performance is evaluated in Section 4. Section 5 provides a summary of the paper.

2. PROBLEM STATEMENT

Microwave breast cancer detection relies on the pulse response between R antennas in a multi-static radar system. There are $M = R(R-1)$ directed antenna pairs. At one time instant, one antenna transmits an ultra-short pulse into the breast, and another antenna records the backscattered signal. The received signal contains the possible backscatter from the tumour, as well as the unwanted incident pulse and reflections

978-1-4673-6997-8/15/$31.00 ©2015 IEEE

from the skin. A scan is complete when signals have been recorded for all antenna pairs.

Assume that we have access to a set of K labelled training scans $Z_{1:K}$ from different breasts, as well as a set of T test scans $Z_{K+1:K+T}$. A scan $Z_k = [z_k^1, z_k^2, \dots, z_k^M]^T$ contains the received pulses from all antenna pairs, where z_k^m denotes the pulse from antenna pair m in the k-th scan. A label $y(Z_k) = -1$ indicates that there is no tumour in scan k, and $y(Z_k) = +1$ indicates the existence of the tumour. The problem is then to classify the test data based on the information obtained from the training data. Our goal is to minimize the miss probability P_M of the system, subject to the constraint that the false positive rate P_F is less than a specified value α.

In our application, we denote the classification results to the test data set $Z_{K+1:K+T}$ by $\hat{y}(Z_{K+1:K+T})$, and the classifier is trained on the training data set $Z_{1:K}$. We define $t_+ = \{t : y(Z_{K+t}) = +1\}$ and $t_- = \{t : y(Z_{K+t}) = -1\}$, and denote the cardinality of each set by T_+ and T_-, respectively. The empirical false positive rate $\hat{P}_F$ is then defined as

$$\hat{P}_F = \frac{\sum_{t \in t_-} (\hat{y}(Z_t) = +1)}{T_-} \; . \tag{1}$$

Similarly, the empirical miss probability $\hat{P}_M$ is

$$\hat{P}_M = \frac{\sum_{t \in t_+} (\hat{y}(Z_t) = -1)}{T_+} \; . \tag{2}$$

Due to the high variation inherent in the empirical false positive rate $\hat{P}_F$, we can accept $\hat{P}_F$ to be larger than α in practice. A scalar performance measure $\hat{e}$ is proposed in [15],

$$\hat{e} = \frac{1}{\alpha} \max\{\hat{P}_F - \alpha, 0\} + \hat{P}_M \; , \tag{3}$$

which serves as the parameter selection criterion in the training stage and the evaluation measure for different classifiers.

3. COST-SENSITIVE ENSEMBLE CLASSIFIER

Our cost-sensitive ensemble classifier consists of three main components: feature extraction, classification, and fusion.

3.1. Feature extraction

The breast scan data $Z_{1:K}$ lie on a space with dimension $\mathbb{R}^{N \times M}$ ($N = 2048$ and $M = 240$ in our system). Classification performed directly in such a high dimensional space will be difficult. We can reduce the dimension by first extracting pertinent features from the received signals. A natural approach is to apply dimension reduction to the individual signals recorded by each antenna pair. During training, PCA is performed on $z_{1:K}^m$ to obtain the principal component coefficients and scores for antenna pair m. We keep the first d scores $x_{1:K}^m$, and use the obtained coefficients to compute the scores for the test data, $x_{K+1:K+T}^m$. These feature vectors are then used as the input to the classifier.

3.2. 2ν-SVM classifier

Support vector machines [16] are among the most effective methods for binary classification. Given a set of K labeled training samples $(x_k, y_k)_{k=1}^K$, where x_k is a feature vector of dimension d, and the label y_k indicates the class of x_k, an SVM first transforms the d-dimensional input vector x_k into a higher dimensional space through a mapping function $h(x)$, in the hope that the transformed data will be easier to classify. The kernel function computes the similarity of the data in the higher dimensional space without computing the coordinates of the data in that space. One popular choice is the radial basis function kernel, parameterized by γ:

$$K(x, x') = \langle h(x), h(x') \rangle = \exp(-\gamma \|x - x'\|^2) \tag{4}$$

The SVM then identifies a hyperplane that has the largest distance to the nearest training data points of any class in the higher dimensional space. These nearest data points are called *support vectors*. The distance between the decision boundary and the support vectors is called the *margin*. The *score function* of the SVM is defined as

$$f(x) = w^T h(x) + b \tag{5}$$

where w is the normal vector to the max-margin hyperplane, and the bias term b defines the decision boundary.

In the ν-SVM [17], the maximum margin solution can be translated into a quadratic programming problem, which penalizes the margin errors by introducing slack variables ϵ_k:

$$\min_{w,b,\epsilon,\rho} \frac{1}{2}\|w\|^2 - \nu\rho + \frac{1}{K}\sum_{k=1}^{K} \epsilon_k \tag{6}$$
$$\text{subject to } \epsilon_k \geq 0, \rho \geq 0, y_k f(x_k) \geq \rho - \epsilon_k, \forall k$$

$\nu \in [0, 1]$ serves as an upper bound on the fraction of margin errors and a lower bound on the fraction of support vectors; and ρ influences the width of the margin.

To allow the assignment of the different costs to different types of errors, the 2ν-SVM [14] was proposed as an extension. The 2ν-SVM takes the form [18]:

$$\min_{w,b,\epsilon,\rho} \frac{1}{2}\|w\|^2 - \nu\rho + \frac{w_+}{K}\sum_{k \in k_+} \epsilon_k + \frac{1 - w_+}{K}\sum_{k \in k_-} \epsilon_k \tag{7}$$
$$\text{subject to } \epsilon_k \geq 0, \rho \geq 0, y_k f(x_k) \geq \rho - \epsilon_k, \forall k.$$

ν and w_+ can be formulated using ν_+ and ν_-:

$$\nu = \frac{2\nu_+\nu_- K_+ K_-}{(\nu_+ K_+ + \nu_- K_-)K} \tag{8}$$

$$w_+ = \frac{\nu_- K_-}{\nu_+ K_+ + \nu_- K_-} = \frac{\nu K}{2\nu_+ K_+} \tag{9}$$

where $\nu_+ \in [0, 1]$ and $\nu_- \in [0, 1]$ bound the fractions of margin errors and support vectors from each class. We can assign different costs to different types of errors by adjusting (ν_+, ν_-).

We apply a cross-validation procedure to choose ν_+ and ν_- so that $\hat{e}$ for a given α is minimized. $\bar{K}$-fold cross validation partitions the training set into $\bar{K}$ folds. The model is trained on all but the $\bar{k}$-th fold, and is tested on the $\bar{k}$-th fold. We iterate through the process until all folds are used as the testing data just once. The empirical NP measures obtained from each fold are then averaged to generate $\hat{e}$.

3.3. Classification architecture

3.3.1. Feature fusion classification approaches

We can concatenate the scores $x^m_{1:K}$ and $x^m_{K+1:K+T}$ from different antenna pairs together to obtain feature vectors $X_{1:K}$ and $X_{K+1:K+T}$. The feature vector X_k is of length Md, which are used by 2ν-SVM classifiers to perform tumour detection. This approach is shown in Figure 1.

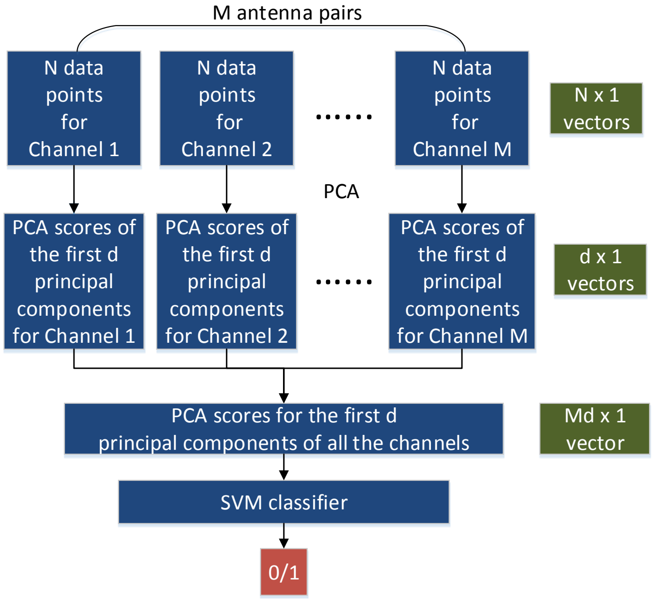

M antenna pairs
N data points for Channel 1
N data points for Channel 2
......
N data points for Channel M
N x 1 vectors
PCA
PCA scores of the first d principal components for Channel 1
PCA scores of the first d principal components for Channel 2
......
PCA scores of the first d principal components for Channel M
d x 1 vectors
PCA scores for the first d principal components of all the channels
Md x 1 vector
SVM classifier
0/1

Fig. 1. The feature fusion classifier approach.

3.3.2. Classifier fusion approach

The concatenated feature vector in the feature fusion approach lies on a high dimensional space $\mathbb{R}^{Md}$. This may lead to poor classification results when there are only limited training data. To address this, we can use the feature vectors from each antenna pair to directly train 2ν-SVM classifiers. The dimension of the feature space is then only d. We average the classifier outputs and apply a threshold to obtain a final decision. The architecture is shown in Figure 2. The threshold η also provides us with a straightforward control over the false positive rate and the miss probability of the ensemble classifier. The value is selected during the cross validation process described in Section 3.2.

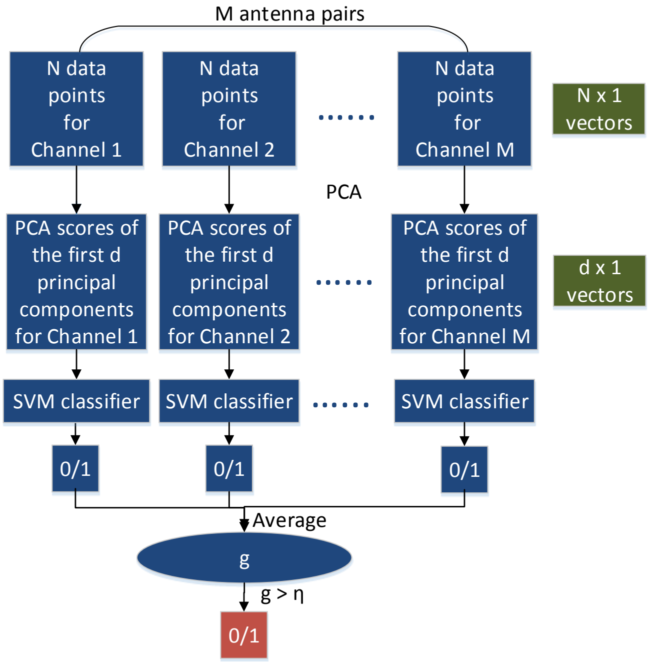
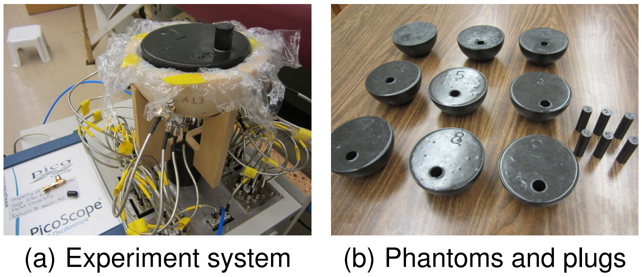

M antenna pairs
N data points for Channel 1
N data points for Channel 2
......
N data points for Channel M
N x 1 vectors
PCA
PCA scores of the first d principal components for Channel 1
PCA scores of the first d principal components for Channel 2
......
PCA scores of the first d principal components for Channel M
d x 1 vectors
SVM classifier
SVM classifier
......
SVM classifier
0/1
0/1
0/1
Average
g
g > η
0/1

Fig. 2. The classifier fusion approach.

4. EXPERIMENT RESULTS

4.1. System overview and data collection

We built a 16-element antenna array time-domain system (Figure 3(a)) and used it to collect data from breast phantoms that are fabricated to mimic the dielectric properties of breast tissues [19]. A short-duration pulse, with spectral content in the 2-4 GHz range, is fed into a 16 × 2 switching matrix which chooses the specific antenna pairs for transmitting and receiving the signal. This process is repeated for each antenna pair, which results in 240 pulses recorded by an oscilloscope with an equivalent-time sampling rate of 200 GSa/s.



(a) Experiment system (b) Phantoms and plugs

Fig. 3. The 9 constructed phantoms and the plugs used to create phantoms with or without tumours.

We constructed 9 breast phantoms with varying dielectric properties. Three are heterogeneous and contain glandular structures that make up approximately 25%, 35%, and 50% of the total volume (Figure 3(b)). We rotate these three phantoms by 120° to mimic 6 new phantoms; we thus have 15 phantoms in total. We can insert a fat plug or a tumour plug to mimic the tumour-free and tumour cases for all phantoms expect Phantom 1, which does not have a plug position so we only have baseline (tumour-free) recordings. We collected 10 sets of baseline scans for each of the 15 phantoms, and 10 sets of tumour scans for each phantom except Phantom 1. Different scans were performed on different days, to mimic the real clinical trial scenario. In all, we have 150 sets of baseline scans and 140 sets of tumour scans. A bandpass filter is applied to eliminate low- and high-frequency noise.

4.2. Performance evaluation

We use 13 breast phantoms to construct the training data, and use the remaining 2 for the test data set. Thus there are $\binom{15}{13} = 105$ training and testing data combinations. We use 13-fold cross validation to identify parameters for each training-testing data combination. We set the number of principal components retained $d = 30$, which gives relatively low NP measures during cross-validation. α is set to 0.05 as the desired upper bound of the false positive rate. We perform parameter selection in two stages: in the first stage, a coarse grid is used; after a "good" region where parameter values lead to relatively low generalization errors is identified, we conduct cross validation on a finer grid (Table 1).

	coarse grid	finer grid for $\alpha = 0.05$
γ	$2^{-15}, 2^{-11}, \ldots, 2^{5}$	$2^{-5}, 2^{-4}, \ldots, 2^{5}$
ν_+	$0.001, 0.01, 0.1, 0.3, 0.6, 1$	$0.001, 0.003, 0.01, 0.03, 0.1$
ν_-	$0.001, 0.01, 0.1, 0.3, 0.6, 1$	$0.001, 0.003, 0.01, 0.03, 0.1$
r	$-0.4, -0.2, \ldots, 0.4$	$-0.3, -0.2, , \ldots, 0.3$

Table 1. Candidate parameter values.

	$\hat{P}_F$	$\hat{P}_M$	average error	$\hat{e}$
Feature fusion	0.019 $[0, 0.05]$	0.017 $[0, 0.1]$	0.019 $[0, 0.05]$	0.093 $[0, 0.1]$
Classifier fusion	0.033 $[0, 0.1]$	0.005 $[0, 0]$	0.021 $[0, 0.05]$	0.233 $[0, 1]$
DMAS	0.053 $[0, 0.13]$	0.946 $[0.85, 1]$	0.546 $[0.53, 0.56]$	1.52 $[1, 2.4]$

Table 2. Average generalization errors and their 10% and 90% quantiles (shown in the square brackets).

Table 2 reports the mean and the 10% and 90% quantiles of the different types of errors across different train-test pairs. We compare performance with two other algorithms. The first is the SVM classifier in [12], which essentially reports the intermediate classifier outputs of Figure 2 as the final result. The average generalization error is 0.129. We also compare to classification based on the maximum voxel intensity of the delay-multiply-and-sum (DMAS) imaging algorithm [6]. We create differential images using the first baseline scan of each phantom as a calibration scan. Figure 4 shows a histogram of the maximum voxel intensities of the generated images.

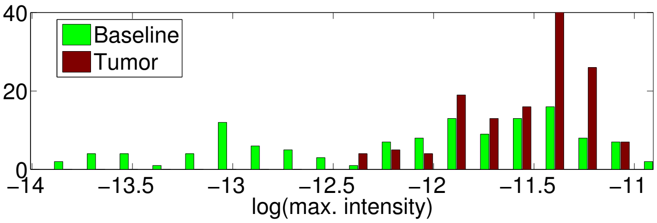


Fig. 4. Maximum image intensity from each scan.

Figure 5 shows the receiving operating characteristic (ROC) curves for the two architectures we have proposed and the DMAS thresholding approach. We observe from this figure and Table 2 that both fusion approaches greatly improve the classification performance. They also control the false positive rate to be less than the desired level $\alpha = 0.05$ for most of the time, leading to a much smaller empirical generalization NP performance measure $\hat{e}$ compared to the DMAS-based detection algorithm. By varying α, we can control the trade-off between the false positive rate and miss probability. On the basis of these results, we cannot conclude that feature fusion or classifier fusion is preferable, and more extensive testing is required.

5. CONCLUSIONS

This paper introduces two different classification architectures for microwave breast cancer detection. The architectures fuse information from all signals recorded by a multistatic antenna array and employ 2ν-SVM classifiers to control the trade-off between the false positive rate and the detection power. Experimental results with measurements collected using a 16-element antenna array prototype applied to tissue-mimicking breast phantoms demonstrate the effectiveness of the proposed architectures.

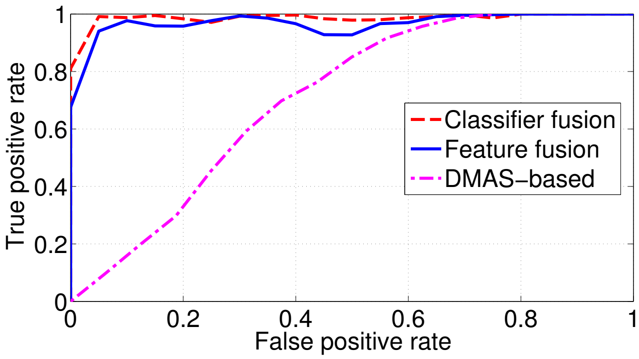


Fig. 5. ROC curves obtained by varying α.

6. REFERENCES

[1] A. Karellas and S. Vedantham, "Breast cancer imaging: a perspective for the next decade," *Med. Phys.*, vol. 35, no. 11, pp. 4878–4897, Nov. 2008.

[2] N.K. Nikolova, "Microwave imaging for breast cancer," *IEEE Microwave*, vol. 12, no. 7, pp. 78–94, Dec. 2011.

[3] L. Sha, E.R. Ward, and B. Stroy, "A review of dielectric properties of normal and malignant breast tissue," in *Proc. IEEE SoutheastCon*, Salt Lake City, UT, May 2002, pp. 457–462.

[4] Tomasz M. Grzegorczyk, P. M. Meaney, P. A. Kaufman, R. M. di Florio-Alexander, and K. D. Paulsen, "Fast 3-d tomographic microwave imaging for breast cancer detection," *IEEE Trans. Med. Imag.*, vol. 31, pp. 1584–1592, Aug. 2012.

[5] E. C. Fear, Xu Li, S. C. Hagness, and M. A. Stuchly, "Confocal microwave imaging for breast cancer detection: localization of tumors in three dimensions," *IEEE Trans. Biomed. Eng.*, vol. 49, pp. 812–822, Aug. 2002.

[6] H. B. Lim, N. T. T. Nhung, E.-P. Li, and N. D. Thang, "Confocal microwave imaging for breast cancer detection: Delay-multiply-and-sum image reconstruction algorithm," *IEEE Trans. Biomed. Eng.*, vol. 55, pp. 1697–1704, June 2008.

[7] E. J Bond, X. Li, S. C Hagness, and B. D Van Veen, "Microwave imaging via space-time beamforming for early detection of breast cancer," *IEEE Trans. Antennas Propagat.*, vol. 51, pp. 1690–1705, Aug. 2003.

[8] M. O'Halloran, E. Jones, and M. Glavin, "Quasi-multistatic MIST beamforming for the early detection of breast cancer," *IEEE Trans. Biomed. Eng.*, vol. 57, pp. 830–840, Apr. 2010.

[9] B. Guo, Y. Wang, J. Li, P. Stoica, and R. Wu, "Microwave imaging via adaptive beamforming methods for breast cancer detection," *J. Electromagn. Waves and Appl.*, vol. 20, no. 1, pp. 53–63, 2006.

[10] Y. Xie, B. Guo, L. Xu, J. Li, and P. Stoica, "Multistatic adaptive microwave imaging for early breast cancer detection," *IEEE Trans. Biomed. Eng.*, vol. 53, pp. 1647–1657, Aug. 2006.

[11] S. K Davis, H. Tandradinata, S. C Hagness, and B. D Van Veen, "Ultrawideband microwave breast cancer detection: a detection-theoretic approach using the generalized likelihood ratio test," *IEEE Trans. Biomed. Eng.*, vol. 52, pp. 1237–1250, July 2005.

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

[13] A. Santorelli, Y. Li, E. Porter, M. Popovic, and M. Coates, "Investigation of classification algorithms for a prototype microwave breast cancer monitor," in *Proc. European Conf. Antennas and Propag. (EuCAP)*, The Hague, The Netherlands, Apr. 2014, pp. 320–324.

[14] H.-G. Chew, R. E. Bogner, and C.-C. Lim, "Dual ν-support vector machine with error rate and training size biasing," in *Proc. Int. Conf. Acoustics, Speech and Signal Proc. (ICASSP)*, Salt Lake City, UT, May 2001, pp. 1269–1272.

[15] C. Scott, "Performance measures for neyman-pearson classification," *IEEE Trans. Inf. Theory*, vol. 53, pp. 2852–2863, Aug. 2007.

[16] B. Schölkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, MIT Press, Cambridge, MA, 2001.

[17] B. Schölkopf, A. J. Smola, R. C. Williamson, and P. L. Bartlett, "New support vector algorithms," *Neural Comput.*, vol. 12, no. 5, pp. 1207–1245, May 2000.

[18] M.A. Davenport, "The 2ν-svm: A cost-sensitive extension of the ν-svm," Tech. Rep. TREE 0504, Dept. of Elec. and Comp. Engineering, Rice University, Houston, TX, Dec. 2005.

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