# Microwave Breast Cancer Detection via Cost-sensitive Ensemble Classifiers: Phantom and Patient Investigation

Yunpeng Li*, Emily Porter, Adam Santorelli, Milica Popović, Mark Coates

*Dept. of Electrical and Computer Engineering, McGill University, Montréal, Québec, Canada*

**Abstract**

Microwave breast screening has been proposed as a complementary modality to the current standard of X-ray mammography. In this work, we design three ensemble classification structures that fuse information from multiple sensors to detect abnormalities in the breast. A principled Neyman-Pearson approach is developed to allow control of the trade-off between false positive rate and the false negative rate. We evaluate performance using data derived from measurements of heterogeneous breast phantoms. We also use data collected in a clinical trial that monitored 12 healthy patients monthly over an eight-month period. In order to assess the efficacy of the proposed algorithms we model scans of breasts with malignant lesions by artificially adding simulated tumour responses to existing scans of healthy volunteers. Tumour responses are constructed based on measured properties of breast tissues and real breast measurements, thus the simulation model takes into account the heterogeneity of the breast tissue. The algorithms we present take advantage of breast scans from other patients or tissue-mimicking breast phantoms to learn about breast content and what constitutes a "tumour-free" and "tumour-bearing" set of measurements. We demonstrate that the ensemble selection-based algorithm, which constructs an ensemble of the most informative classifiers, significantly outperforms other detection techniques for the clinical trial data set.

*Keywords:*
Microwave breast cancer detection, sensor fusion, biomedical monitoring, clinical trial, ensemble selection.

*Corresponding author. Address: Room 633, McConnell Engineering Building, 3480 University Street, Montreal, Québec, Canada H3A 0E9.

*Email addresses:* yunpeng.li@mail.mcgill.ca (Yunpeng Li), emily.porter@mail.mcgill.ca (Emily Porter), adam.santorelli@mail.mcgill.ca (Adam Santorelli), milica.popovich@mcgill.ca (Milica Popović), mark.coates@mcgill.ca (Mark Coates)

## 1. Introduction

Early detection of breast cancer is vital for successful treatment [1]. Microwave imaging and detection methods have been intensely researched in recent years as a complementary modality for breast cancer screening. Such methods are based on the reported inherent contrast of dielectric properties of healthy and malignant breast tissues over the microwave frequency range [2, 3]. Microwave techniques promise non-invasive screening with low-cost system fabrication and operation and have been applied to other fields including stroke detection [4]. Scans do not require breast compression and can be repeated frequently since no ionizing radiation is used. The aim is not to replace mammography, ultrasound, or MRI, but to develop an alternative approach that can act as an early warning system to flag the need for more comprehensive testing.

Most of the previous work on microwave breast cancer screening has concentrated on imaging. Algorithms are applied to measurement data to generate images that can be interpreted by a clinical expert. Microwave radar and microwave tomography are two common techniques in the microwave imaging field. Tomographic methods are used to reconstruct a dielectric profile of breast tissues [5] by solving an ill-conditioned inverse problem. Radar methods, on the other hand, generate a map of scattering regions within the breast. Tomography methods have been applied in experimental imaging of both phantoms [6, 7] and patients [8, 5]. Radar imaging approaches include beamforming algorithms [9, 10, 11] and hypothesis testing techniques [12]. Results have been reported for delay-and-sum and other beamforming algorithms on data collected in clinical trials [13, 14, 11].

Recently, some research has explored the application of machine learning techniques, in particular classifiers, to measurements collected from microwave breast cancer screening systems [15, 16, 17, 18, 19]. Classification techniques have been applied to characterize a tumour using microwave backscatter [15, 16] with the assumption that the tumour has already been detected. In [15], architectural tissue features such as shape and size are inferred from the backscatter by using linear classifiers with local discriminant bases and principal component analysis (PCA). Conceição et al. introduce a support vector machine (SVM)-based classifier that distinguishes between benign and malignant tumours according to their shape [16].

There has also been some recent work towards the detection task for microwave breast cancer screening systems. In [17], a discrete cosine transform (DCT) is applied to the received signal for feature extraction, and neural networks are used to detect the tumour existence, size and location. In [18], Byrne et al. apply SVM to features extracted from backscattered signals using PCA. A separate SVM classifier is applied to each measured signal; a tumour is detected for the breast if the majority of the classifiers decided that it

was present. Building on this work, we described a strategy in [19] that fused data from random antenna pairs to improve the SVM classifier accuracy.

Our work focuses on the development of a microwave breast cancer screening system and the associated algorithms that can process measurements to make a decision as to whether a tumour is present in the breast. This system could offer women the option of self-screening at home on a regular (e.g., monthly) basis. The monthly (as opposed to annual) tests would be especially beneficial to those in the high-risk category, as frequent monitoring increases the chance of early-stage tumour diagnosis and, consequently, successful treatment. We envision that the system will track breast health by comparing the current breast scan to past scans of the same patient and to other patient scans, stored in a clinical database.

With this long-term goal and motivation in mind, several essential milestones have been reached to date. We have developed a time-domain microwave radar system for breast screening. Time-domain measurements potentially offer advantages over frequency-domain, including faster scan times and more cost-effective equipment solutions, with the drawback of a slightly lower signal-to-noise ratio [6]. We have demonstrated successful imaging of tumours in realistic tissue phantoms [20]. Recently, we have conducted a clinical trial with 12 patient volunteers for breast health monitoring [21].

This paper presents a novel application of classification methods to clinical data collected from a microwave breast screening system to detect the presence of a tumour. Our main contributions compared to state-of-the-art work in this domain are the following: (i) we employ a principled Neyman-Pearson approach to select algorithmic parameters in order to control the false positive rate while minimizing the false negative rate (most past work in microwave breast cancer detection did not differentiate between these two types of errors); (ii) we design three ensemble classification architectures to fuse information from different antenna pairs; (iii) we demonstrate the performance of our classification techniques using data collected in a clinical trial that monitored patients monthly over an eight-month period. Preliminary results concerning this work and the efficacy of imaging-based algorithms with clinical trial data have been published in abbreviated forms [22, 23]. This paper extends our previous work by proposing the ensemble selection-based classification method which significantly outperforms existing methods. It also provides a more detailed description of our algorithms, a new data-adaptive tumour response simulation procedure that factors in the heterogeneous propagation environment inside the breast, and a more complete performance evaluation involving both a breast phantom data set and a clinical trial data set.

The remainder of the paper is organized as follows: Section 2 introduces our system, data, and the ensemble classifier. We report and discuss experiment results in Section 3 and Section 4, and provide a summary in Section 5.

## 2. Materials and Methods

### *2.1. System overview*

The system uses multiple antenna sensors to collect the transmitted and reflected signals from the breast. The core of the system (Figure 1) is a hollowed-out hemispherical dielectric radome, which houses both the breast under test and the 16-element antenna array. The radome is a ceramic dielectric made from alumina (with relative permittivity $\epsilon_r = 9.6$) [24]. The antennas are travelling-wave, resistively-loaded sensors that are designed for operation in the vicinity of breast tissues [25]. When a breast scan recording begins, a short-duration Gaussian-modulated pulse is generated and shaped, using a passive microwave filter, such that its frequency content is concentrated in the 2-4GHz range [26]. The pulse is amplified and then input into an automated $16 \times 2$ switching matrix that selects each antenna as the transmitter in turn. The pulse is scattered off of the breast tissues, i.e., at all interfaces between tissue types, and is then collected by the selected receiving antenna. An equivalent-time sampling oscilloscope records the data. Then, a different transmit-receive antenna pair is selected until all possible combinations have been cycled through. With $16$ antennas, a total of $240$ signals are obtained per breast scan.

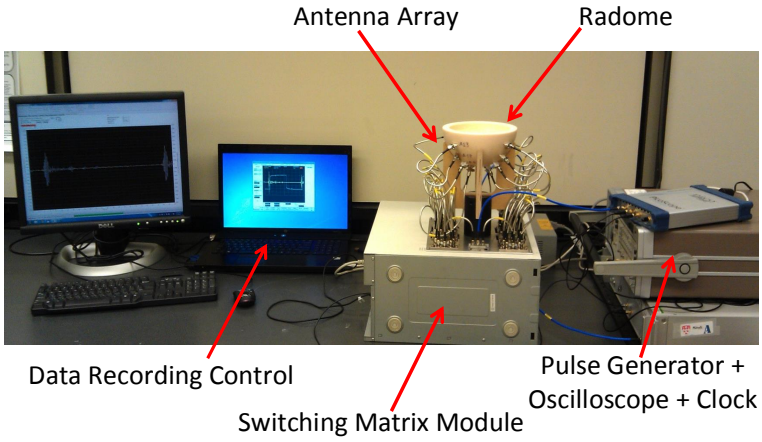
For performing breast scans on patients, the system is integrated in a way that it can be easily used to collect patient measurements in clinical trials. All equipment is placed under a table and the patient lies facing down on the table with their breast in the radome, which protrudes through an aperture in the table. This setup allows for a comfortable breast scan. As the radome is designed for the largest possible breast size, a gel or liquid is required to fill air gaps between the skin and the radome walls. We use ultrasound gel because it conforms easily when under light pressure and is approved for medical applications [21]. The gel, with relative permittivity $\epsilon_r = 68$ and conductivity $\sigma = 3$ S/m at the centre frequency $3$ GHz, also provides a lossy background such that multiple reflections from the skin or radome walls are attenuated. The total duration of a scan is less than 2 minutes.

### *2.2. Data*

We collected data with both tissue-mimicking breast phantoms and volunteers to evaluate proposed algorithms.

#### *2.2.1. Breast phantom data collection*

We constructed $9$ breast phantoms with varying dielectric properties. The breast phantoms are created from a mixture of polyurethane, graphite, and carbon black. We constructed four different mixtures. These are designed to mimic the dielectric properties of skin, gland, fat, and tumour. The description of the construction process of the phantoms



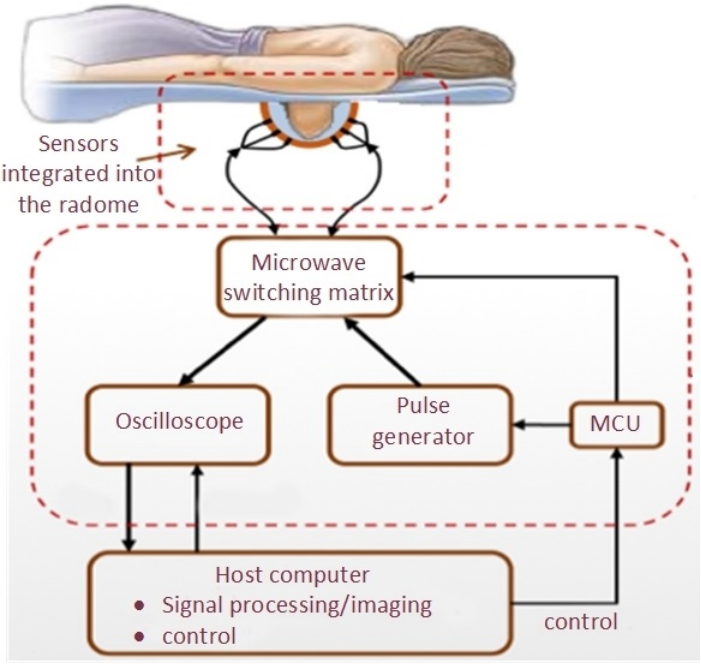




Figure 1: Top: the experiment system we use to collect the data. Bottom: a graphical illustration of the system prototype for this experiment.

and also the properties of the used materials are given in [27]. We adapted the procedure described in [28], adding acetone as a thinning agent to improve the mixability and increase the overall permittivity of the samples.

Among these phantoms, three are heterogeneous. They contain a mixture of glandular- and fat-mimicking solids, surrounded by a thin ($2.5$ mm) layer of the skin-mimicking material. The three phantoms have varying proportions of glandular structures (approximately $25\%$, $35\%$, and $50\%$ of the total volume, respectively). The other six phantoms are comprised of a fat-mimicking mixture surrounded by skin-mimicking material. Note that even in these "homogeneous" phantoms, the nature of the mixing process means that there is still substantial variation in the dielectric properties in different parts of the phantom. In eight of the phantoms, We can insert a fat plug to mimic the tumour-free cases and we can insert a tumour plug to mimic the tumour-bearing cases, respectively. A tumour plug contains a kernel of tumour-mimicking mixture, of approximately spherical shape and $1$ cm in diameter. It is found that malignant tumours are more irregular shaped and benign tumours would be more round and regular shaped [29]. The spherical-shaped tumor is not fully realistic, however it is a well-accepted design for simulation and early experiments [12, 30]. The phantoms and plugs are shown in Figure 2.

The heterogeneous phantoms can be rotated to present the measurement system with a different structure. We rotated the three heterogeneous phantoms by $120^\circ$ and $240^\circ$ to mimic $6$ new phantoms, as the rotation changes the relative distances between the antenna array and breast structure. Although the rotated phantoms are not completely independent, the rotation changes the paths presented to all antenna pairs. We thus have $15$ phantoms in total. We collected $10$ sets of baseline scans for each of the $15$ phantoms, and $10$ sets of tumour scans for each phantom except Phantom $1$ (no plug). Different scans were performed on different days, to mimic a real clinical trial scenario. In all, we have $150$ sets of baseline scans and $140$ sets of tumour scans. The sample length is $N = 2048$ and the sampling rate is $200$ GHz.

*2.2.2. Clinical trial*

In addition to the breast phantom data, we also performed breast scans on $12$ healthy volunteers using our radar system. The clinical trial lasted $8$ months, and involved $48$ patient visits, with each volunteer visiting approximately once per month. The patient volunteers visited a minimum of two and a maximum of six times. The volunteers ranged in age from $21$ to $77$, with bra cup sizes from A to D. Data were recorded by an oscilloscope with a sampling rate of $40$ GHz and a signal length of $1024$ samples. Table 1 shows the number of visits obtained from each of the $12$ volunteers.

We collected measurements of the left breast and the right breast from the same person at each visit. Thus, there are $48 \times 2 = 96$ measurements collected. Since we only have

Figure 2: The breast phantoms and plugs constructed to collect the phantom measurements in controlled experiments.

Table 1: Number of visits for each volunteer.

| Volunteer index | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Number of visits | 3 | 3 | 4 | 5 | 2 | 6 | 6 | 4 | 4 | 4 | 3 | 4 | 48 |

clinical data from healthy volunteers, there are no tumour-bearing measurements in our original data set. However, we can simulate the tumour responses for each volunteer, based on the transmitted pulses from the antennas and the dielectric properties of breast tissue and tumour.

For a given volunteer, we randomly choose half of the visits as nominal visits, i.e., measurements from those visits form the tumour-free data set for that volunteer. Measurements from the other half of the visits are injected with artificial tumour responses, and we call these visits the "tumour-bearing" visits. We end up with $48$ nominal measurements and $48$ tumour-bearing measurements in total.

To simulate the tumour response, we need to first calculate the antenna positions of the $16$-element antenna array system. The antennas are located on a hemisphere defined by $x^2/a^2 + y^2/b^2 + z^2/c^2 = 1$, where $z \geq 0$ is the depth information, $a = b = c = 7.3$ cm. The antenna positions match those in our multistatic radar system prototype. The breast is modelled as a smaller hemisphere, with $a' = b' = c' = 7$ cm. This geometry (Figure 3) is chosen to approximate that of the physical system from which the clinical data were collected.

In practice, the antennas will be immersed in a medium which will couple the signals

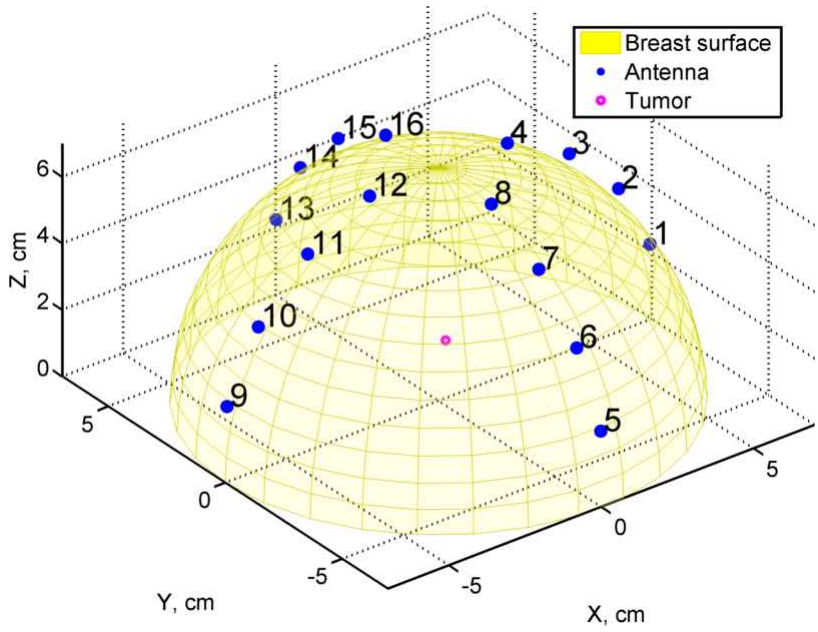



Figure 3: The model used to construct tumour responses. The breast surface is approximated as a hemisphere. Positions of antennas A1 to A16 (blue) correspond to the locations in our experimental system. In this example, we position a tumour (pink) at $(1, 1, 1)$ cm.

optimally for their propagation into the breast. In our phantom experiments, the radome, which houses the antennas, has been fabricated from a ceramic material which represents this matching medium. It is lossless and its relative permittivity has been selected to be that of the average reported value for the human breast tissue across the frequency range of interest (~1 - 6 GHz).

To generate a tumour response for one antenna pair, let's first denote the measured received signal from that antenna pair by $r(t)$ with a corresponding frequency representation $R(\omega)$.

$$R(\omega) = G(\omega)H(\omega)\,, \tag{1}$$

where $G(\omega)$ is the frequency domain representation of the transmitted signal, $H(\omega)$ is the response of the healthy breast in our measurement system [31].

When a tumour exists at location $p_0$, the frequency domain representation $R^t(p_0, \omega)$ of the tumour response $r^t(p_0, \omega)$ is given by

$$\begin{aligned} R^t(p_0, \omega) &= \Gamma G(\omega) H_t(p_0, \omega) \\ &= \Gamma R(\omega) H_t(p_0, \omega)/H(\omega) \end{aligned} \tag{2}$$

where the superscript $t$ indicates the tumour response, $H_t(p_0, \omega)$ is the response of the tumour-bearing breast, and the factor $\Gamma$ introduces additional attenuation to the tumour response. If $\Gamma = 1$, the channel model assumes that the entire signal is reflected at the interface with the tumour. In practice the dielectric contrast between the tumour and the surrounding tissue is limited, so part of the signal will not be reflected, which we account for by setting $\Gamma$ to a smaller value, e.g. 0.5 as in Section 3.

The simulated tumour response $R^t(p_0, \omega)$ is a function of the received pulse $R(\omega)$, which experiences attenuation and delay through the real breast environment. Thus, the heterogeneity of the breast tissue has been accounted for in the tumour response model.

Denote by $\widetilde{H}(\omega)$ and $\widetilde{H}_t(p_0, \omega)$ theoretical models for the responses of the healthy breast and the tumour, respectively.

$$\widetilde{H}(\omega) = e^{-j(k_{im} d^d_{im} + k_{br} d^d_{br})}\ . \tag{3}$$

$d^d_{im}$ and $d^d_{br}$ are lengths of the direct path for antenna pair $m$, in the immersion medium and the breast tissue, respectively. The superscript $d$ denotes the direct path. The wavenumber for the immersion medium $k_{im}(\omega) = 2\pi/\lambda_{im}(\omega) = \sqrt{\epsilon_{im}(\omega)}\omega/c$, where $\lambda_{im}(\omega)$ is the wavelength at frequency $\omega$ in the immersion medium and $c$ is the speed of light. The wavenumber for the breast tissue $k_{br} = 2\pi/\lambda_{br}(\omega) = \sqrt{\epsilon_{br}(\omega)}\omega/c$ , where $\epsilon_{br}$ is the average breast tissue complex relative permittivity (see (8)). Similarly,

$$\widetilde{H}_t(p_0, \omega) = e^{-j(k_{im}(\omega) d^t_{im} + k_{br}(\omega) d^t_{br})}\ . \tag{4}$$

$d_{im}^t$ and $d_{br}^t$ are the lengths of the shortest path between the antenna pair through the tumour position $p_0$, in the immersion medium and the breast tissue, respectively.

We model $H(\omega)$ by

$$H(\omega) = \widetilde{H}(\omega)S(\omega)\,, \tag{5}$$

where $S(\omega)$ captures all of the aspects of the response from the healthy breast not adequately described by the Debye model. And we approximate $H_t(p_0, \omega)$ by assuming that the tumour response has a common $S(\omega)$ which capture all effects not described by the Debye model with the healthy breast

$$H_t(p_0, \omega) = \widetilde{H}_t(p_0, \omega)S(\omega). \tag{6}$$

.

By combining equations (2)-(6), we have the complete tumour response generation model

$$R^t(p_0, \omega) = \Gamma R(\omega) e^{-j(k_{im}(d_{im}^t - d_{im}^d) + k_{br}(d_{br}^t - d_{br}^d))}\ . \tag{7}$$

For different breasts, the received signals will experience different delays and attenuations, due to different breast sizes and shapes, as well as different dielectric properties of the breasts. Since the simulated tumour response is a function of $R(\omega)$, which is the frequency content of the received signals, as shown in Equation (7), the simulated tumour responses are different due to variation in sizes, shapes, and dielectric properties of real breasts.

The average breast tissue complex relative permittivity $\epsilon_{br}(\omega)$ used to calculate $k_{br}$ varies with $\omega$ and we adopt the Debye model [32]:

$$\epsilon_{br}(\omega) = \epsilon_\infty + \frac{\Delta\epsilon}{1 + j\omega\tau} + \frac{\sigma_s}{j\omega\epsilon_0}\ . \tag{8}$$

$\epsilon_0 = 8.854 \times 10^{-12}$ F/m is the permittivity of free space, and the other four model parameters – the dielectric constant of the material at infinite frequency $\epsilon_\infty$, $\Delta\epsilon = \epsilon_s - \epsilon_\infty$ where $\epsilon_s$ is the static dielectric constant, the pole relaxation constant $\tau$, and the static conductivity $\sigma_s$ are chosen to approximate the dielectric properties of breast tissue, as described below.

We estimated average relative permittivity values for the patients who participated in our clinical trial and observed a range of $[25, 40]$ at 3 GHz. We use the data provided in [2] as a starting point for identification of suitable Debye parameters. Lazebnik et al. conducted permittivity measurements of excised tissues, and identified four tissue groups. The first three groups corresponded to different ratios of adipose (fatty) tissue to fibroconnective or glandular tissue. The fourth group corresponded to malignant tissue. In [33], Lazebnik et al. fit Debye parameters for each group using the median measured values.

Based on our range of estimated average relative permittivity values at 3 GHz from clinical trial data [21], the most appropriate models are the group 2 ($31\%$-$84\%$ adipose tissue) and group 3 ($85\%$-$100\%$ adipose tissue) models.

We specify suitable ranges for each of the four Debye model parameters. First, we set $\Delta\epsilon \in [20, 32.08]$. The upper value corresponds to the group 2 model from [33], and the lower value ensures that $\epsilon_{br} \geq 25$ at 3 GHz. The value of $\epsilon_\infty$ has noticeable impact on the complex permittivity only above $4$ GHz, which is beyond the operational regime for our system. We therefore set it to $5.57$, the group 2 value. The data reported in [2] and [3] provide an indication of the expected broadband conductivity for a heterogeneous breast composed of both adipose and glandular tissue. For the observed range of permittivities, we expect a range of 1–2 S/m at 3 GHz. With this range in mind, we choose $\sigma_s \in [0.36, 0.52]$ S/m and $\tau \in [8.68, 13]$ ps. The upper bound of $\sigma_s$ and the lower bound of $\tau$ match the group 2 values; the other bounds are selected to ensure the conductivity $\sigma \geq 1$ S/m at 3 GHz.

To generate tumour signals for each patient, we drew values uniformly at random from the identified ranges of the Debye parameters. The generated electrical permittivity and conductivity curves are plotted in Figure 4. Different $\epsilon_{br}(\omega)$ values produce different wavenumbers $k_{br}$ used in (7), which lead to tumour responses with different time delay (determined by the real parts of $k_{im}$ and $k_{br}$) and different attenuation (determined by the imaginary parts of $k_{im}$ and $k_{br}$).

The final part in the generation of the tumour responses is the tumour position. For each patient volunteer, we randomly selected a location in the upper outer quadrant of the breast model, where the majority of breast cancer tumours are located [34, 35, 36]. One sample nominal measurement and tumour-bearing measurement of Volunteer 1 are shown in Figure 5.

We can also simulate tumour response for breast phantoms and verify the simulation model with the tumour response recovered from real tumour-bearing measurements, which indicates measurements collected with a tumour plug. The recovered tumour response is obtained by subtracting the tumour-bearing measurement from an aligned calibration measurement, where the aligned measurement is the first measurement collected from the same phantom with a fat plug, to be used as the reference measurement. Figure 6 shows one sample recovered tumour response for A1A4, and the simulated tumour response with attenuation factor $\Gamma = 0.5$, which is shifted to achieve the maximal correlation with the recovered tumour response. We can see that the simulated tumour response is a reasonable approximation of the recovered tumour response in this example (a close match is not expected near the end of the signals where noise tends to dominate the recovered response).

For the clinical trial data, a calculation of the range of tumour response delay based on the electrical permittivity at the central frequency 3 GHz and breast model geometry

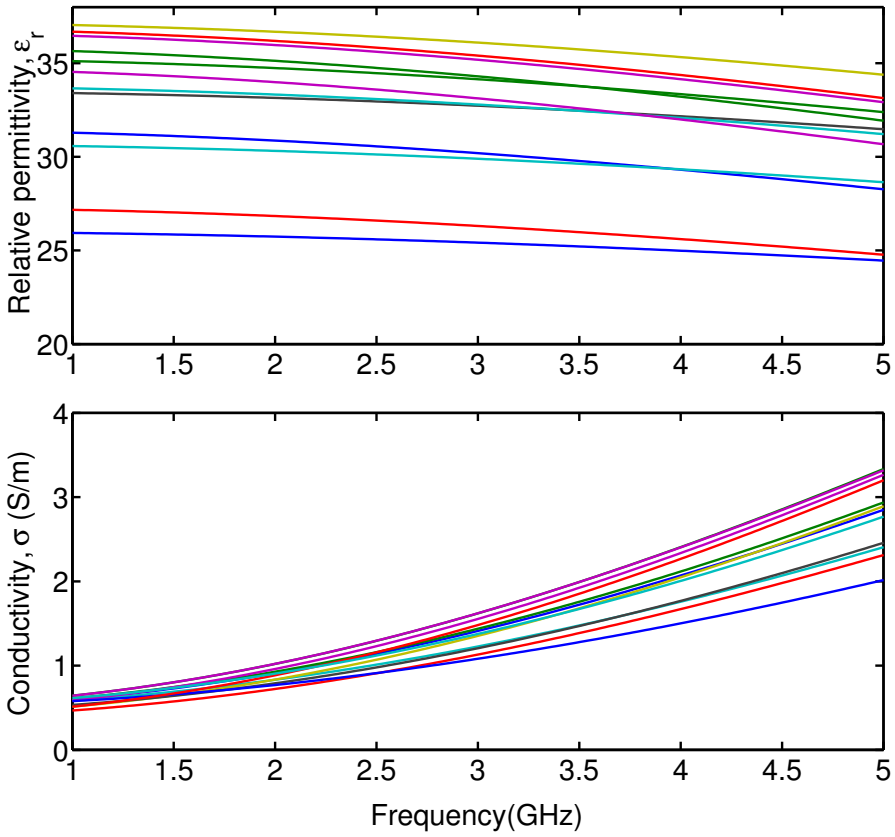



Figure 4: The relative permittivity and conductivity curves generated using the Debye model, as used for generation of the tumour signals based on (3)-(8). Each curve corresponds to a different patient volunteer. Different curves arise because the Debye model parameters are drawn uniformly at random from specified ranges.

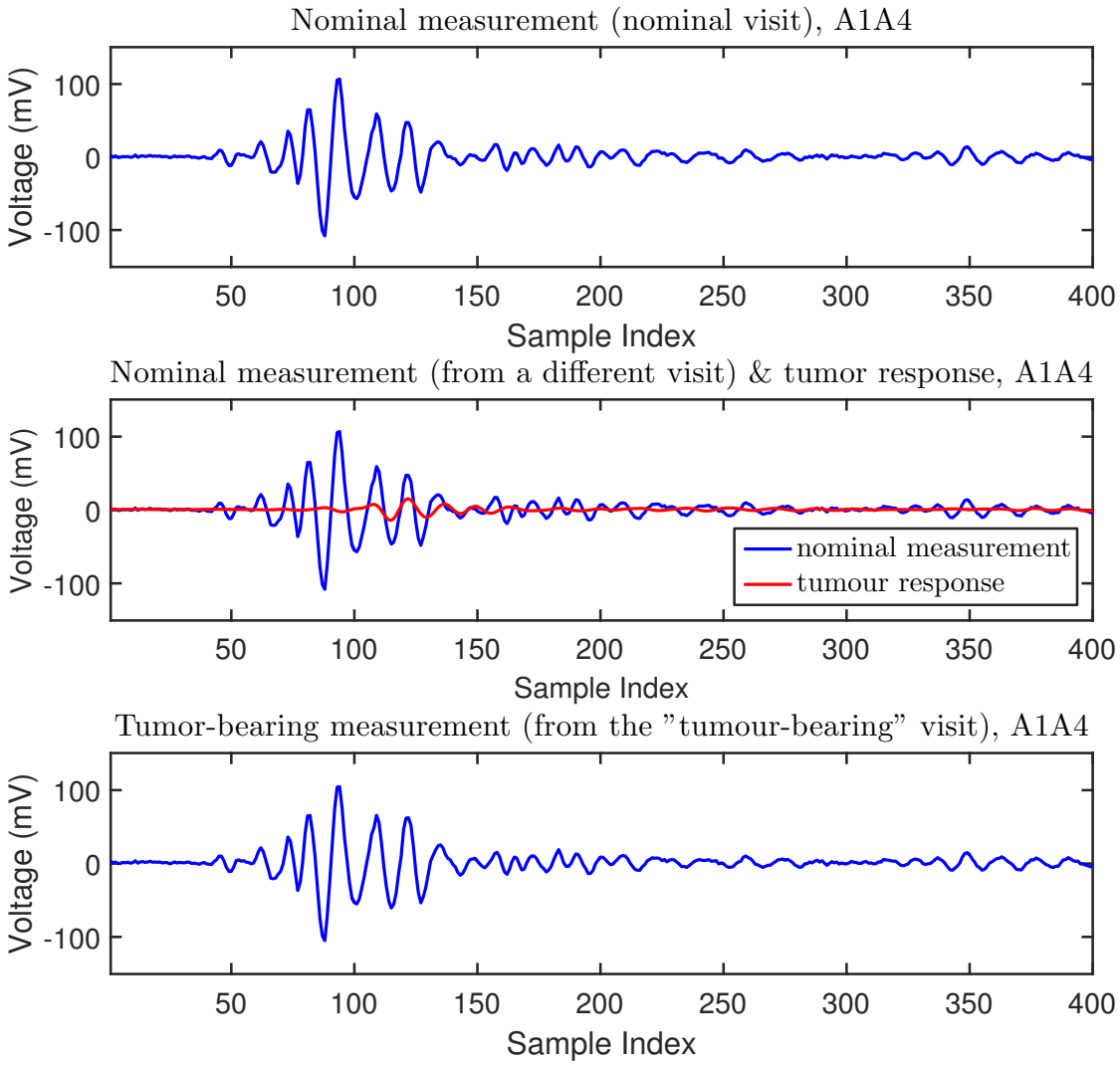



Figure 5: Top: The A1A4 signal from the first nominal visit of Volunteer 1. Middle: The A1A4 nominal signal from the first "tumour-bearing" visit of Volunteer 1 and the generated tumour response with the attenuation factor $\Gamma = 1$. Bottom: The artificial tumour-bearing signal generated by adding the two signals from the middle plot, i.e., adding the tumour response to the nominal measurement from the tumour-bearing visit. The sampling rate is 40 GSa/s.

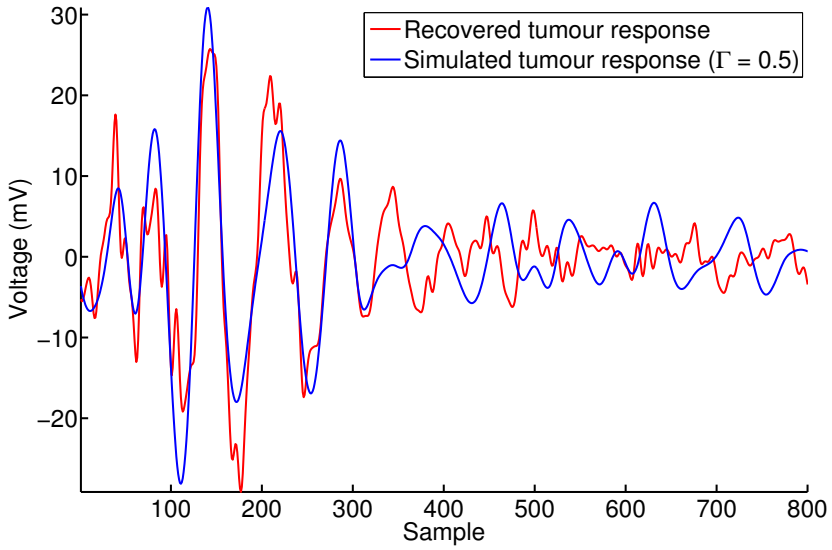



Figure 6: Red: the A1A4 tumour response, recovered from tumour-bearing measurement from Phantom 7. Blue: the simulated tumour response generated for Phantom 7 with $\Gamma = 0.5$, shifted by 39 samples to obtain the maximal correlation with the recovered tumour response. They are windowed to display only the possible tumour response range. The sampling rate is 200 GHz.

reveals that all tumour response peak amplitudes should span no more than $200$ samples for the $40$ GHz sampling rate. Moreover, tumour response delays larger than $140$ samples are rare and only happen for a few tumour locations and a small subset of antenna pairs. To reduce interference caused by fluctuations of measurements among different visits, we window clinical trial scans to concentrate on the region where tumour responses are most likely to appear, using measurements between the $61^{\text{st}}$ and $200^{\text{th}}$ sample for feature extraction. The starting value 61 samples is an empirical value that are used to roughly align data from different scans. Since the system has undergone updates during the 8 months of scans, scans between some volunteer visits can have offsets. Data are thus first roughly aligned by applying a window staring from 60 samples before the peak amplitude of the antenna pair A1A2 measurement. Since A1 and A2 are one of the closest antenna pairs, we do not expect any tumour responses to happen before the arrival of the received measurement of A1A2. Thus, we apply a window [61, 200] to the data. A detailed description of calculation steps for the window is provided in the appendix.

### 2.3. Classification methods

#### 2.3.1. Cost-sensitive classification

Suppose we have collected $K$ labelled training measurements, which we denote by $Z_{1:K}$. These different measurements come from different patients taken on different scan occasions. We also obtained $T$ test measurements $Z_{K+1:K+T}$. A single measurement $Z_k = [z_k^1, z_k^2, \ldots, z_k^{\tilde{M}}]^{\mathrm{T}}$ includes the received signals from all antenna pairs, where $z_k^m$ is the $\tilde{N}$-sample signal measured by antenna pair $m$ during the $k$-th scan. In a practical setting, labels could be assigned by diagnosticians, using the microwave scans in conjunction with other breast cancer screening techniques such as mammography and ultrasound. A positive label indicates that there is no tumour in scan $k$, and a negative label indicates the existence of a tumour. Our task is to assign a label to each test measurement.

We aim to minimize the false negative rate $P_M$ of the system, subject to the constraint that the false positive rate $P_F$ is less than a specified value $\alpha$. This is because although minimizing the false negative rate reduces the chance of missed early detection, it is also important to prevent overdiagnosis and overtreatment of the breast cancer [37]. When training a classifier, we try to control the false positive rate and bound its value. Since we can only calculate the empirical false positive rate and the empirical false negative rate, these empirical rates can exhibit high variation around the true value unless the training and testing sets are large, so we should not automatically eliminate classifiers that exhibit $\widehat{P}_F$ larger than $\alpha$. Scott et al. propose a scalar performance measure $\hat{e}$ in [38] that can be used to gauge the performance of a classifier that is required to control the false positive rate to lie below $\alpha$:

$$\hat{e} = \frac{1}{\alpha} \max\{\hat{P}_F - \alpha, 0\} + \hat{P}_M \ . \tag{9}$$

We use this metric as the parameter selection criterion in the training stage and for evaluation of the proposed classifier.

The three primary components of our cost-sensitive ensemble classifier are feature extraction, classification, and fusion. We describe how we address each of these tasks as follows.

#### 2.3.2. Feature extraction

The $R$-antenna multistatic radar system transmits an ultrawideband pulse from each antenna and measures the response at every other antenna. It thus collects measurements from $\tilde{M} = R(R-1)$ directed antenna pairs. Each received signal contains the unwanted direct pulse and skin reflections, as well as any possible tumour responses. Low-power signals can be recorded when the transmitting and receiving antenna are far apart from each other, e.g. on opposite sides of the breast. These signals can be very weak due

to attenuation, as well as varying significantly between visits. Through data inspection, we observe that the signal variations are considerably large with respect to its amplitude. Thus, incorporating these into the processing is effectively just adding noise and can degrade classification performance. We therefore discard the signals from an antenna pair if the median peak amplitude across all training data is less than a threshold (see Section 3 for the choice of the threshold value).

The unprocessed breast scan data reside in a high-dimensional space ($\mathbb{R}^{N\times M}$). Here $N$ is the window length (see Section 2.2.2) and $M$ is the number of retained directed antenna pairs. If a classifier is applied directly to data in a high dimensional space, training is challenging and the performance is likely to be very poor. The standard approach is to extract features that capture the key information embedded in the received signals. This achieves substantial reduction in the dimension of the training data that is passed to the classifier.

The signals collected by different antenna pairs vary greatly, but our experiments indicate that the baseline signals for a specific pair have similar content for different scans and different patients. We therefore apply *principal component analysis (PCA)* individually to the ensemble of training signals for each antenna pair (Figure 7(a)). Principal component analysis successively identifies orthogonal components ($N \times 1$ vectors). The variability of the different signals comprising the data set is maximized in the direction of the first component. Successive components also maximize the variability, but they must satisfy the constraint of being orthogonal to all earlier components. For each measured signal, there is a *score* associated with each component; this corresponds to the projection of the signal onto the component. We denote the vector of $d$ scores by $x_k^m$ for antenna pair $m$ and scan $k$. The first score $x_{k,1}^m$ corresponds to the projection of $z_k^m$ onto the first principal component $v_1^m$, i.e., $x_{k,1}^m = \langle z_k^m, v_1^m \rangle$, where $\langle \cdot, \cdot \rangle$ denotes the dot-product.

In training our classifier, we only retain the first $d$ scores associated with each measurement (see Section 3 for a discussion on the choice of $d$). By doing so, we hope to eliminate from consideration all of the signal elements that are common to all measurements arising from a specific antenna pair, whether there is a tumour present or not. These elements include the direct pulse between the antennas and artefacts generated by reflections at the skin interface.

The inputs to the classifier for training purposes are the score vectors $x_k^m$ for $k = 1, \ldots, K$ and $m = 1, \ldots, M$. The labels for the test data are derived by applying the classifier to the score vectors $x_k^m$ for $k = K+1, \ldots, K+T$. These vectors are derived by projecting the test data onto the principal components identified from the training data.

*2.3.3. $2\nu$-SVM classifier*

Support vector machines [39] have proven to be a very effective method for binary classification. Given a set of $K$ labelled 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 $h()$. The mapping aims to make the transformed data easier to classify with a decision boundary defined by $f(x) = w^T h(x) + b = 0$, where $w$ is the normal vector to the separating hyperplane, and $b$ is the bias term of the decision boundary.

In general the two classes cannot be separated, so slack variables $\epsilon_k \geq 0$ need to be introduced. A value $\epsilon_k > 0$ indicates that the $k$-th data element lies on the wrong side of the decision boundary. To allow assignment of different costs to different types of errors, Chew et al. introduced the $2\nu$-SVM with the following objective function [40, 41]:

$$\min_{w,b,\epsilon,\rho} \frac{1}{2}||w||^2 - \nu\rho + \frac{w_+}{K}\sum_{k\in k_+}\epsilon_k + \frac{1-w_+}{K}\sum_{k\in k_-}\epsilon_k \tag{10}$$
$$\text{subject to } \epsilon_k \geq 0, \rho \geq 0, y_k f(x_k) \geq \rho - \epsilon_k, \forall k \ .$$

Here $k_+$ denotes the set of data elements with $y_k = +1$, and $k_-$ denotes the set of data elements with $y_k = -1$. We can have different penalties for margin errors depending on whether the data label is positive or negative, and the parameter $w_+$ controls the relative weight of the penalties. We can express the problem in a different way by introducing parameters $\nu_+ \in [0, 1]$ and $\nu_- \in [0, 1]$ (hence the name $2\nu$-SVM). These parameters bound the fractions of margin errors and support vectors from each class. Using these parameters we can replace $\nu$ and $w_+$:

$$\nu = \frac{2\nu_+\nu_-K_+K_-}{(\nu_+K_+ + \nu_-K_-)K} \tag{11}$$

$$w_+ = \frac{\nu_-K_-}{\nu_+K_+ + \nu_-K_-} = \frac{\nu K}{2\nu_+K_+} \ , \tag{12}$$

where $K_+$ and $K_-$ are the numbers of positively- and negatively- labelled data, respectively. We can assign different costs to different types of errors by adjusting $(\nu_+, \nu_-)$.

In training a classifier, our aim is to minimize $\hat{e}$ from (9) for a specified bound $\alpha$ on the false positive rate. We apply a cross validation procedure to choose $\nu_+$, $\nu_-$, and $\gamma$. The $\bar{K}$-fold cross validation procedure partitions the training set into $\bar{K}$ *folds* (disjoint groups). For each candidate parameter set, the model is trained on all but the $\bar{k}$-th fold, and is then tested on the $\bar{k}$-th fold. We iterate through this process until every fold has been used once as the testing data. The empirical Neyman-Pearson measures obtained from each fold are then averaged to generate $\hat{e}$ for each candidate choice of parameters. We select the parameter values that minimize $\hat{e}$.

*2.3.4. Fusion architecture*

The feature extraction procedure outlined above generates $M$ vectors (one for each retained antenna pair) for each scan, with each vector containing $d$ elements. We must choose how to use these vectors in a classification architecture.

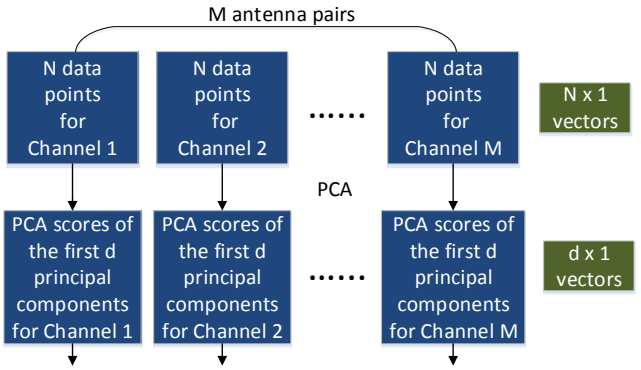
**Feature fusion approach** A simple approach is to combine all of the score vectors associated with a single scan into a large vector with $Md$ elements. This is then the feature vector associated with that scan and can be used as an element in a single $2\nu$-SVM classifier. This approach is illustrated in Figure 7(b).

Although simple, this approach has some drawbacks; even if we use a relatively small $d$, we have a feature vector in $\mathbb{R}^{Md}$. For example, if we retain $10$ scores per antenna pair and keep measurements from every antenna pair, we still have an input vector containing $2400$ elements for our system. The dimension has been reduced dramatically (from $1024 \times 240 = 245760$) but it is still classification in a high-dimensional feature space. This may lead to poor classification results when there are limited training data.

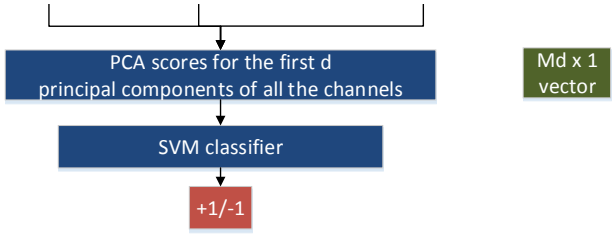
**Classifier fusion approach** An alternative approach is to use feature vectors from each antenna pair to directly train multiple $2\nu$-SVM classifiers. The dimension of the feature space for each classifier is then only $d$. We average classifier outputs and apply a threshold to obtain a final decision. The architecture is shown in Figure 7(c). The threshold $\eta$ also provides us with a straightforward control over the false positive rate and the false negative rate of the ensemble classifier. We use common $\nu_+$, $\nu_-$, and $\gamma$ for all of the classifiers, and select these parameter values and $\eta$ during the cross validation process described in Section 2.3.3.

**Ensemble selection approach** We may expect that measurements from a subset of antenna pairs contain more useful information than others, due to the antenna configuration and non-uniform tumour location distribution [34]. Instead of combining classification results from all antenna pairs as in the classifier fusion approach, we utilize an ensemble selection scheme to form an ensemble of more *informative* classifiers. Ensemble selection [42] has been shown for a variety of data sets to outperform most other ensemble learning techniques including stacking, bagged decision trees, and boosted decision trees.

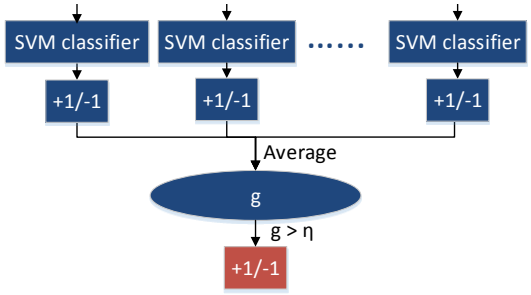
The ensemble selection algorithm first builds a model "library" by training different base models. Each base model is a $2\nu$-SVMs trained using the data from a single antenna pair. But many base models are trained using each antenna pair by varying the hyperparameters used by the $2\nu$-SVM. The number of base models can thus be



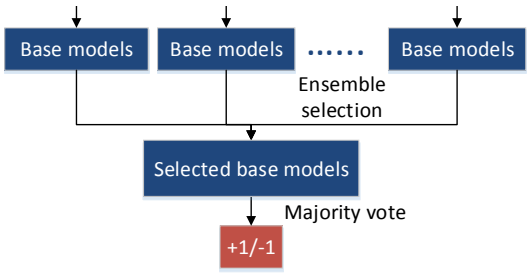

(a) Feature extraction.




(b) Feature fusion.




(c) Classifier fusion.




(d) Ensemble selection.

Figure 7: Different ensemble classifiers (b)-(d) all follow the feature extraction step in (a).

large, due to the large number of choices of hyper-parameters (see Table 2) and the relatively large number of antenna pairs.

When we have a small data set, we can only form a small validation set, which can lead to overfitting. A cross-validated base model is proposed in [42] to make use of all training data available. It consists of $\bar{K}$ "sibling" models trained on $\bar{K}$ cross validation folds with the same model parameters. In the training stage, a cross-validated model's performance is the average of the performance of its $\bar{K}$ sibling models. In the prediction stage, a cross-validated model's final decision is the average of the $\bar{K}$ outputs from all siblings.

The algorithm then selects $Q$ cross-validated base models from the model library with the smallest values of Neyman-Pearson measures $\hat{e}$ in the training stage. The final classification output is a majority vote of outputs from those selected base models. The architecture is presented in Figure 7(d).

## 3. Results

For the clinical data set, we use all of the data (nominal and tumour) from one volunteer as the test data, and use all of the measurements from the other volunteers as the training data. Since there are 12 volunteers, we have 12 training-testing pairs for the clinical data set. We construct 15 training-testing pairs for the 15 phantoms following the same approach. The median peak amplitude threshold for the antenna pair selection introduced in the beginning of Section 2.3.2 is set to $20$ mV, as we observe through data inspection that direct pulses from the same antenna pair with peak amplitudes less than $20$ mV can be highly distorted and vary significantly among different scans for the same patient. Figures generated for data inspection are not included in the paper, for conciseness and keeping the focus of the paper on the detection methodology. As a result, 185 of the 240 antenna pairs are retained for the clinical trial set and 212 antenna pairs are retained for the breast phantom data set. The number of principal components retained is set to $d = 30$. This value was chosen based on earlier experimentation with breast phantoms [22], and we do not observe significant performance variation between candidate values $d \in \{30, 50\}$. The number of best base models retained in ensemble selection, is set to $Q = 100$, as again no notable performance differences are found among a set of candidate values $Q \in \{50, 75, 100, 150, 200\}$. We perform parameter selection using cross validation over values specified in Table 2 for other parameters in the feature fusion and classifier fusion approaches. The SVM hyper-parameters are $\gamma, \nu_+, \nu_-$. As shown in Table 2, there are 11 candidate values for $\gamma$, 18 for $\nu_+$ and 18 for $\nu_-$. Thus, there are $11 \times 18 \times 18 = 3564$ different combinations of SVM hyper-parameters. For ensemble selection, the 3564 different SVM hyper-parameters are used to produce a model library

consisting of $3564 \times 185 = 659340$ base models for classification with the clinical trial data set, where $185$ is the number of antenna pairs retained. Constructing many base models ensures that we have a large pool from which to choose those that have high cross validation accuracy. The number of folds of cross validation $\bar{K}$ is set to the number of volunteers or phantoms in the training set. So $\bar{K} = 14$ for the breast phantom data set and $\bar{K} = 11$ for the clinical trial data set. Each fold contains data from one volunteer or one breast phantom.

Table 2: Candidate parameter values used in the ensemble classifier.

| | |
|---|---|
| $\gamma$ | $2^{-15}, 2^{-13}, \ldots, 2^{5}$ |
| $\nu_+$ | $1 \times 10^{-5}, 3 \times 10^{-5}, 1 \times 10^{-4}, 3 \times 10^{-4},$ $0.001, 0.003, 0.01, 0.03, 0.1, 0.2, 0.3, 0.4, \ldots, 1$ |
| $\nu_-$ | $1 \times 10^{-5}, 3 \times 10^{-5}, 1 \times 10^{-4}, 3 \times 10^{-4},$ $0.001, 0.003, 0.01, 0.03, 0.1, 0.2, 0.3, 0.4, \ldots, 1$ |
| $r$ | $-0.4, -0.3, \ldots, 0.4$ |

Table 3 reports the mean and the $10\%$ and $90\%$ quantiles of different types of errors across different training-testing pairs for the three proposed architectures. For breast phantom data, we report detection performance evaluated with the original 9 phantoms as introduced in Section 2.2.1, as well as the 15-phantom data set which is obtained by rotating the heterogeneous phantoms. We also compare ensemble classifiers with classification algorithms based on two imaging algorithms: delay-multiply-and-sum (DMAS) [9] and the generalized likelihood ratio test (GLRT) [12]. To generate the images, we consider the first measurement of each volunteer/phantom as a calibration measurement. Differential measurements are then created using the later scans (subtracting one signal from the other after time-alignment). The maximum voxel intensity in the generated DMAS or GLRT image is used as the classifier input, as proposed in [23]. We record the input pulse from the microwave system for the GLRT algorithm to generate signal templates. The Debye parameters or the permittivity values used in the imaging algorithms are based on the estimated permittivity values from the clinical data or the phantom measurements [21].

Figure 8 to Figure 10 show the receiver operating characteristics (ROCs) for different classification algorithms for the phantom and clinical data sets. Since we observe in Table 3 that classification results with 9 phantoms and 15 phantoms are similar, we only plot ROCs with 9 phantoms for the breast phantom data set. Different operating points of the ROC are obtained by adjusting the desired false positive rate upper bound $\alpha$ in the training stage. This leads to many different classifiers with different parameters; the depicted ROC curves are then vertical averages of the operating points of these classifiers, calculated over

Table 3: Average performance of the classification approaches (with $10\%$ and $90\%$ quantiles in square brackets). The columns show the estimated the false positive rate, the false negative rate, the average error, and the Neyman-Pearson error measure (Equation (9)), as evaluated on the test data. False positive rate target $\alpha$ is set to $0.1$. Shaded entries indicate the smallest Neyman-Pearson error for each data set.

| Data | Classifier | $\hat{P}_F$ | $\hat{P}_M$ | average error | $\hat{e}$ |
|---|---|---|---|---|---|
| Phantom data (9 phantoms) | feature fusion | 0.01 [0.00, 0.06] | 0.02 [0.00, 0.12] | 0.02 [0.00, 0.09] | 0.02 [0.00, 0.12] |
| | classifier fusion | 0.01 [0.00, 0.06] | 0.04 [0.00, 0.16] | 0.03 [0, 0.08] | 0.04 [0.00, 0.16] |
| | ensemble selection | 0.01 [0.00, 0.06] | 0.03 [0.00, 0.10] | 0.02 [0.00, 0.08] | 0.03 [0.00, 0.10] |
| | DMAS | 0.01 [0.00, 0.07] | 0.89 [0.40, 1.00] | 0.45 [0.20, 0.53] | 0.90 [0.40, 1.07] |
| | GLRT | 0.01 [0.00, 0.07] | 0.89 [0.40, 1.00] | 0.46 [0.27, 0.50] | 0.90 [0.47, 1.00] |
| Phantom data (15 phantoms) | feature fusion | 0.01 [0.00, 0.10] | 0.03 [0.00, 0.00] | 0.02 [0.00, 0.10] | 0.03 [0.00, 0.00] |
| | classifier fusion | 0.02 [0.00, 0.10] | 0.01 [0.00, 0.00] | 0.01 [0, 0.05] | 0.01 [0.00, 0.00] |
| | ensemble selection | 0.09 [0.10, 0.10] | 0.00 [0.00, 0.00] | 0.05 [0.05, 0.05] | 0.00 [0.00, 0.00] |
| | DMAS | 0.06 [0.00, 0.11] | 0.90 [0.89, 1.00] | 0.48 [0.44, 0.56] | 1.02 [0.89, 1.11] |
| | GLRT | 0.01 [0.00, 0.00] | 0.92 [0.78, 1.00] | 0.47 [0.50, 0.50] | 1.00 [1.00, 1.00] |
| Clinical data $\Gamma = 1$ | feature fusion | 0.13 [0.00, 0.48] | 0.78 [0.41, 1.00] | 0.46 [0.25, 0.61] | 1.75 [0.50, 4.14] |
| | classifier fusion | 0.13 [0.00, 0.55] | 0.76 [0.35, 1.00] | 0.44 [0.18, 0.67] | 1.76 [0.35, 5.05] |
| | ensemble selection | 0.08 [0.00, 0.43] | 0.39 [0.00, 1.00] | 0.24 [0.00, 0.50] | 1.06 [0.00, 3.67] |
| | DMAS | 0.08 [0.00, 0.30] | 0.78 [0.18, 1.00] | 0.47 [0.12, 0.83] | 1.53 [0.18, 3.70] |
| | GLRT | 0.04 [0.00, 0.15] | 0.92 [0.70, 1.00] | 0.51 [0.18, 0.75] | 1.25 [0.70, 2.20] |
| Clinical data $\Gamma = 0.5$ | feature fusion | 0.14 [0.00, 0.77] | 0.96 [0.0.76, 1.00] | 0.55 [0.47, 0.83] | 2.18 [0.94, 7.57] |
| | classifier fusion | 0.11 [0.00, 0.35] | 0.89 [0.50, 1.00] | 0.50 [0.25, 0.64] | 1.65 [0.50, 3.47] |
| | ensemble selection | 0.06 [0.00, 0.29] | 0.88 [0.47, 1.00] | 0.47 [0.32, 0.50] | 1.30 [0.90, 2.33] |
| | DMAS | 0.08 [0.00, 0.30] | 0.92 [1.00, 1.00] | 0.57 [0.31, 0.83] | 1.75 [1.00, 3.70] |
| | GLRT | 0.04 [0.00, 0.15] | 0.92 [0.70, 1.00] | 0.51 [0.18, 0.75] | 1.25 [0.70, 2.20] |

all the different training-testing pairs. Since the values in Table 3 correspond to a single set of classification parameters determined by the Neyman-Pearson learning framework for $\alpha = 0.1$, they do not necessarily lie on the average ROC curves. The figures include for comparison the ROCs obtained for the thresholded imaging methods, calculated by varying the threshold and averaging the performance over different subsets of the data. Each subset of the data is created using a leave-8-out approach, which includes all but 8 measurements.

We observe from Table 3 and Figure 8 that all three ensemble classifiers exhibit very good performance when applied to the phantom data. The imaging-and-thresholding algorithms, especially the one based on GLRT, perform much worse than the classifiers. This is probably because the algorithmic assumptions of the imaging algorithms are poorly matched to the phantom data. For the clinical data, when the attenuation factor $\Gamma = 1$, indicating that we adopt tumour responses simulated from the propagation model and real measurement, the ensemble selection-based algorithm has a clear performance advantage over the other algorithms, as shown in Table 3 and Figure 9. When $\alpha = 0.1$, the ensemble classifier achieves an average false positive rate of $0.08$ and a false negative rate of $0.39$. When $\Gamma = 0.5$, which means that we impose further attenuation that may be unaccounted for in the propagation model, all classifiers' performance degrades due to the lower signal-to-noise ratio, with the ensemble selection-based classification algorithm still being the best one, as shown in Figure 10.

## 4. Discussion

We now discuss the proposed ensemble classifiers for cost-sensitive tumour detection in the context of the obtained results. The low signal-to-noise ratio nature of the tumour response requires a relatively large number of antennas in the microwave system. We investigated the effectiveness of fusion of information from these antenna pairs using several architectures (Figure 7), including the fusion of features from different antenna pairs (the feature fusion approach), the fusion of base classifiers from all antenna pairs (the classifier fusion approach), and the fusion of the most informative base classifiers (the ensemble selection approach). We demonstrate through Table 3 and Figure 8–10 that the ensemble selection structure is the most effective.

The cost-sensitive aspect of the classifier is realized primarily through the $2\nu$-SVM. The algorithms control the trade-off between the false positive rate and the false negative rate by selecting algorithmic parameters to minimize a scalar Neyman-Pearson measure (Equation 9). Classification results show that the ensemble classifiers are able to control the two types of errors successfully, approximately restricting the false positive rate below a specified maximum value.

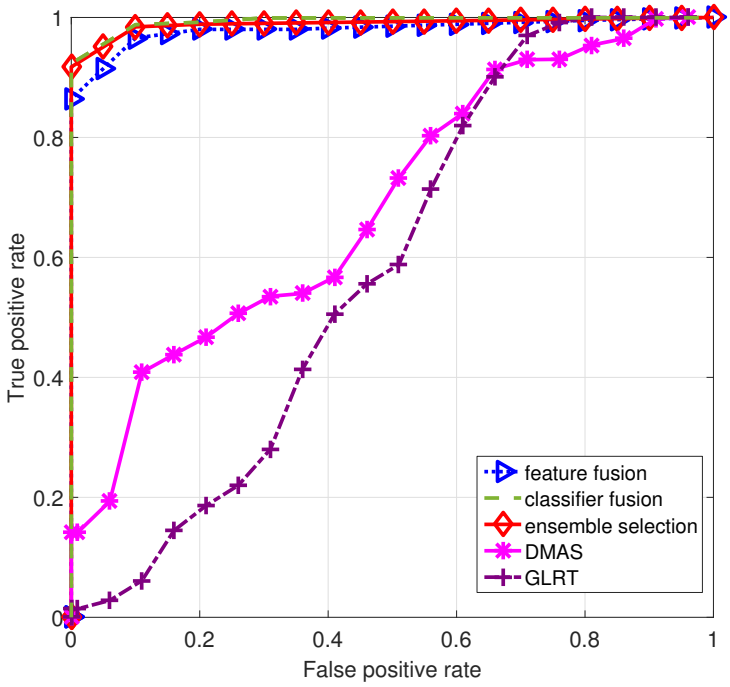



Figure 8: ROC curves of different algorithms for the phantom data (9 phantoms).

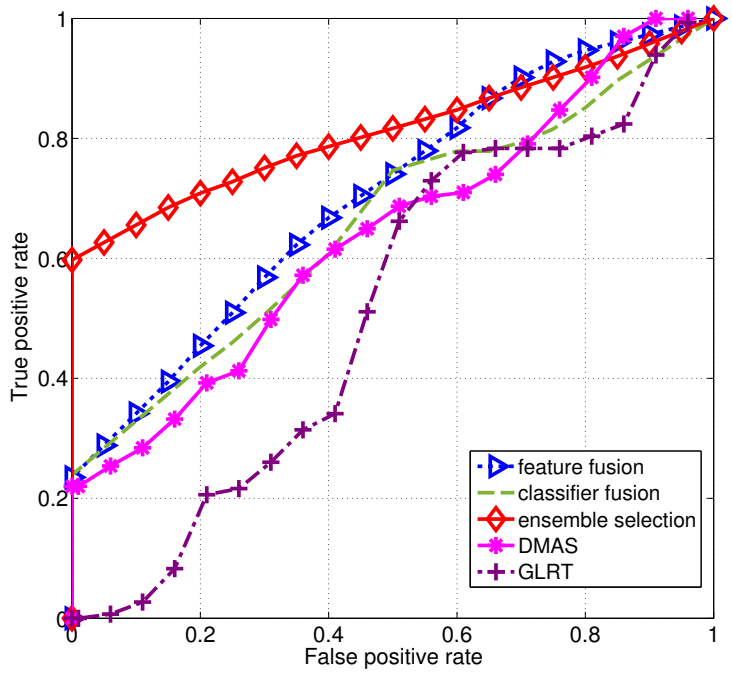



Figure 9: ROC curves of different algorithms for the clinical data ($\Gamma = 1$).

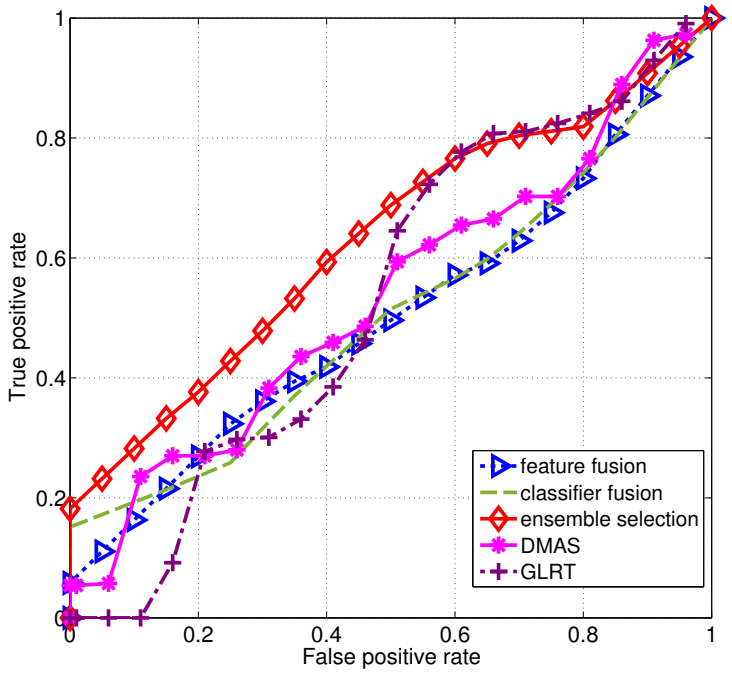



Figure 10: ROC curves of different algorithms for the clinical data ($\Gamma = 0.5$).

We compare the proposed algorithms with traditional imaging-based algorithms. Imaging algorithms are model-based approaches because they incorporate tumour response propagation models. They also require calibration data or baseline measurements to enable extraction of the tumour responses from the raw signals during pre-processing. The calibration process is difficult due to skin reflections, alignment accuracy, and other interference. The GLRT algorithm imposes stricter model assumptions than DMAS, and the detection results in Figure 8 indicate that it leads to worse performance due to the mismatch between the model and the phantom data. In contrast, ensemble classifiers perform classification via a data-centric approach. They learn features from both tumour-free and tumour-bearing measurements in the training data set, and construct a decision boundary based on those features. No tumour response propagation models are required.

To evaluate the performance of the proposed algorithms, we applied them to two data sets, one generated from breast phantom measurements and the other from clinical measurements combined with numerically-generated tumour responses. Few previous classification algorithms for microwave-based breast cancer screening have been validated with any form of clinical data. Numerical simulations are used to validate the algorithms in [15, 18], and the methods in [16, 17, 19] are assessed with breast phantom data.

To construct the breast phantom data set, we collected only one measurement of each phantom each day. The system was switched off and the phantoms were taken out of the radome after each day's measurements. The data thus includes re-positioning errors, which are likely to be unavoidable in a clinical setting. The re-positioning error, breast heterogeneity, and other measurement issues introduce distortions and unpredictable signal delays. There is thus a mismatch between the models underpinning the imaging algorithms and the data being analyzed. As a result the performance of the imaging algorithms is relatively poor. The ensemble classifiers, on the other hand, exhibit very good detection performance (Figure 8).

For the clinical data set, measurements were obtained from a clinical trial that resulted in scans of healthy patients. In order to use this data to assess the performance of the detection algorithms, we simulate tumour responses and add them to a subset of the measurements. Our approach involves applying filters (frequency-dependent attenuation and delay) to the signals recorded at the antennas. Because we process the clinically-recorded measurements, we automatically incorporate many of the effects of the heterogeneous propagation channel. Experiments with the heterogeneous breast phantoms provide qualitative support to this method of constructing a tumour response (see Figure 6). The results suggest that the detection algorithms find the data based on clinical measurements more challenging than the phantom data. There is more variability in the breast size and structure and there are additional challenges in the measurement procedure. As indicated by the example Figure 5, the simulated tumour responses have a very small amplitude, and

the training dataset is small. It is encouraging to see from Figure 9 and Figure 10 that the proposed ensemble classifiers can detect approximately sixty-five percent of the tumours for a small false-positive rate, but for practical application in a clinical setting, the performance would require significant improvement. There are multiple avenues for achieving this, including hardware improvements to obtain signals with reduced noise, better scanning procedures, larger training sets (including historical scan data from the breast under test), and more sophisticated processing algorithms.

The ensemble selection approach outperforms the other ensemble classifiers and imaging-based algorithms, and this highlights the importance of fusing information from different antenna pairs in an effective way. This is particularly essential when the signal-to-noise ratio is very low, as is the case in the clinical data set. Some antenna pairs generate signals that have limited information content and high levels of noise. Eliminating these antenna pairs from consideration in the ensemble classifier via an ensemble selection procedure leads to improved performance.

## 5. Conclusion

In this paper we presented cost-sensitive ensemble classification techniques for microwave breast screening. We evaluated their performance with two different sets of data, one based on clinical experiments and one based on breast phantom measurements. In the preliminary clinical trial, all participants were healthy, so to assess the detection performance we added simulated tumour responses to some of the scan data. The tumour responses were constructed by processing measured clinical data, and we validated the response construction process using phantom experiments. Although the added tumour is simulated, the background signal corresponds to what is measured in a clinical setting. Thus, the impact of the heterogeneous breast tissue on signal propagation is incorporated into the simulation model. This approach provides an alternative method for algorithmic validation that complements the assessment based on breast phantom.

We described three fusion strategies to perform classification using cost-sensitive support vector machines. By employing cost-sensitive ensemble classification architectures, the algorithms were able to choose thresholds in a principled manner to ensure that the false positive rate lies below a specified maximum value. Among the three strategies we presented, The ensemble selection procedure significantly outperformed the other ensemble classifiers and imaging-based approaches. Classification performance degraded when the algorithms were applied to the clinical data set which exhibits a lower signal-to-noise ratio, with the ensemble selection-based algorithm being the only relatively effective approach. This motivates improvement of the measurement system and procedure, and the further development of classification algorithms. In particular, it is of interest to construct an algorithm that can make more effective use of a patient's own past scans.

**Acknowledgments**

This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC), the Fonds de recherche du Québec - Nature et technologies (FRQNT), and Prompt Québec [grant PJT2011-03].

**Appendix**

The windowing range from the 61st sample to the 200th sample for the clinical trial data range can be calculated based on the electrical permittivity at the central frequency 3GHz. Here we briefly describe the calculation steps.

The speed of our signal is $v = \frac{c}{\sqrt{\epsilon_r}}$, where $c = 3 \times 10^8$ m/s is the speed of light. The average relative permittivity in the tissue is estimated to be in the range $[27, 34]$ at 3 GHz [21]. A larger relative permittivity indicates a lower propagation speed, which should be used to calculate the upper limit of the delay. Thus, we use $\epsilon_r = 34$ for calculation.

Thus,

$$v = \frac{c}{\epsilon_r} = \frac{3 \times 10^8}{\sqrt{34}} = 5.145 \times 10^7 \text{m/s} \tag{13}$$

From the geometry of the antennas which are located in a hemisphere with radius $7.3$ cm, the distance from the transmitting antenna to the receiving antenna via the tumour, is at most $25$ cm. So the maximum travel time is

$$t_1 = \frac{0.25}{v} = 4.859 \times 10^{-9} \text{ s} = 4859 \text{ ps}\,. \tag{14}$$

Since the sampling rate for the clinical data set is $40$ Gsa/s, the sample interval $dt = 25$ ps. This can lead to a tumour response delay $\frac{t_1}{dt} \approx 194$ samples.

Moreover, based on the relative geometry between the antenna pairs and the tumour, tumour responses from most antenna pairs should have much smaller traversal distances. The rare signals with longer distances have very low amplitudes. Thus, we consider applying windowing to include only tumour responses with travel distance less than $18$ cm. So, the maximum travel time $t_2$ is

$$t_2 = \frac{0.18}{v} = 3.50 \times 10^{-9} \text{ s} = 3500 \text{ ps}\,. \tag{15}$$

Since the sample interval $dt = 25$ ps, the maximum tumour response delay of these antenna pairs is $\frac{t_2}{d_t} = 140$ samples. As the start of the window is the $61$st sample as described in the main text, the windowing range is set to between the 61st and 200th sample.

[1] Amer. Cancer Soc., Facts & Figures 2014 (2014).
URL http://www.cancer.org/research/cancerfactsstatistics/

[2] M. Lazebnik, L. McCartney, D. Popovic, C. Watkins, M. Lindstrom, J. Harter, S. Sewall, A. Magliocco, J. Booske, M. Okoniewski, S. Hagness, A large-scale study of the ultrawideband microwave dielectric properties of normal, benign and malignant breast tissues obtained from cancer surgeries, Phys. Med. Biol. 52 (20) (2007) 6093–6115.

[3] T. Sugitani, S. Kubota, S. Kuroki, K. Sogo, K. Arihiro, M. Okada, T. Kadoya, M. Hide, M. Oda, T. Kikkawa, Complex permittivities of breast tumor tissues obtained from cancer surgeries, Applied Physics Lett. 104 (25) (2014) 253702.

[4] M. Persson, A. Fhager, H. Trefna, Y. Yu, T. McKelvey, G. Pegenius, J.-E. Karlsson, M. Elam, Microwave-based stroke diagnosis making global prehospital thrombolytic treatment possible, IEEE Trans. Biomed. Eng 61 (2014) 2806–2817.

[5] T. M. Grzegorczyk, P. M. Meaney, P. A. Kaufman, R. M. di Florio-Alexander, K. D. Paulsen, Fast 3-d tomographic microwave imaging for breast cancer detection, IEEE Trans. Med. Imag. 31 (2012) 1584–1592.

[6] X. Zeng, A. Fhager, P. Linner, M. Persson, H. Zirath, Experimental investigation of the accuracy of an ultrawideband time-domain microwave-tomographic system, IEEE Trans. Instrum. Meas. 60 (2011) 3939–3949.

[7] N. R. Epstein, P. M. Meaney, K. D. Paulsen, 3D parallel-detection microwave tomography for clinical breast imaging, Rev. Sci. Instrum. 85 (12) (2014) 124704.

[8] H. Jiang, C. Li, D. Pearlstone, L. L. Fajardo, Ultrasound-guided microwave imaging of breast cancer: Tissue phantom and pilot clinical experiments, Med. Phys. 32 (8) (2005) 2528–2535.

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

[10] M. O'Halloran, E. Jones, M. Glavin, Quasi-multistatic MIST beamforming for the early detection of breast cancer, IEEE Trans. Biomed. Eng. 57 (2010) 830–840.

[11] D. Byrne, I. Craddock, Time-domain wideband adaptive beamforming for radar breast imaging, IEEE Trans. Antennas Propagat. 63 (2015) 1725–1735.

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

[13] J. Bourqui, J. M. Sill, E. C. Fear, A prototype system for measuring microwave frequency reflections from the breast, Intl. J. Biomed. Imaging 12 (2012) 1–12, Article ID 85123.

[14] M. Klemm, I. Craddock, J. Leendertz, A. Preece, D. Gibbins, M. Shere, R. Benjamin, Clinical trials of a UWB imaging radar for breast cancer, in: Proc. European Conf. Antennas and Propag. (EuCAP), Barcelona, Spain, 2010, pp. 1–4.

[15] S. K. Davis, B. D. Van Veen, S. C. Hagness, F. Kelcz, Breast tumor characterization based on ultrawideband microwave backscatter, IEEE Trans. Biomed. Eng. 55 (2008) 237–246.

[16] R. C. Conceição, H. Medeiros, M. O'Halloran, D. Rodriguez-Herrera, D. Flores-Tapia, S. Pistorius, SVM-based classification of breast tumour phantoms using a UWB radar prototype system, in: Proc. URSI General Assembly and Scientific Symposium (GASS), Beijing, China, 2014, pp. 1–4.

[17] S. A. AlShehri, S. Khatun, A. B. Jantan, R. S. A. R. Abdullah, R. Mahmud, Z. Awang, Experimental breast tumor detection using NN-based UWB imaging, Prog. Electromagn. Res. (PIER) 111 (2011) 447–465.

[18] D. Byrne, M. O'Halloran, E. Jones, M. Glavin, Support vector machine-based ultrawideband breast cancer detection system, J. Electromagn. Waves and Appl. 25 (13) (2011) 1807–1816.

[19] A. Santorelli, Y. Li, E. Porter, M. Popović, 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, 2014, pp. 320–324.

[20] E. Porter, E. Kirshin, A. Santorelli, M. Coates, M. Popović, Time-domain multistatic radar system for microwave breast screening, IEEE Antennas Wireless Propag. Lett. 12 (2013) 229–232.

[21] E. Porter, M. Coates, M. Popovi, An early clinical study of time-domain microwave radar for breast health monitoring, IEEE Trans. Biomed. Eng. 63 (3) (2016) 530–539.

[22] Y. Li, A. Santorelli, O. Laforest, M. Coates, Cost-sensitive ensemble classifiers for microwave breast cancer detection, in: Proc. Intl. Conf. Acoustics, Speech and Signal Proc. (ICASSP), Brisbane, Australia, 2015.

[23] Y. Li, E. Porter, M. Coates, Imaging-based classification algorithms on clinical trial data with injected tumour responses, in: Proc. European Conf. Antennas and Propag. (EuCAP), Lisbon, Portugal, 2015.

[24] E. Porter, E. Kirshin, A. Santorelli, M. Popović, Microwave breast screening in the time-domain: Identification and compensation of measurement-induced uncertainties, Prog. Electromagn. Res. (PIER) 55 (2013) 115–130.

[25] H. Kanj, M. Popović, A novel ultra-compact broadband antenna for microwave breast tumor detection, Prog. Electromagn. Res. 86 (2008) 169–198.

[26] A. Santorelli, M. Chudzik, E. Kirshin, E. Porter, A. Lujambio, I. Arnedo, M. Popović, J. D. Schwartz, Experimental demonstration of pulse shaping for time-domain microwave breast imaging, Prog. Electromagn. Res. 133 (2013) 309–329.

[27] A. Santorelli, O. Laforest, E. Porter, M. Popović, Image classification for a time-domain microwave radar system: Experiments with stable modular breast phantoms, in: European Conf. Antennas and Propag. (EuCAP), Lisbon, Portugal, 2015.

[28] J. Garrett, E. Fear, Stable and flexible materials to mimic the dielectric properties of human soft tissues, IEEE Antennas and Wireless Propag. Lett. 13 (2014) 599–602.

[29] R. M. Rangayyan, N. M. El-Faramawy, J. E. L. Desautels, O. A. Alim, Measures of acutance and shape for classification of breast tumors, IEEE Trans. Med. Imag. 16 (6) (1997) 799–810.

[30] M. Klemm, J. A. Leendertz, D. Gibbins, I. J. Craddock, A. Preece, R. Benjamin, Microwave radar-based breast cancer detection: Imaging in inhomogeneous breast phantoms, IEEE Antennas Wireless Propag. Lett. 8 (2009) 1349–1352.

[31] R. Bracewell, The Fourier transform and its applications, New York, NY.

[32] P. J. W. Debye, Polar molecules, The Chemical Catalog Co., New York, NY, 1929.

[33] M. Lazebnik, M. Okoniewski, J. H. Booske, S. C. Hagness, Highly accurate Debye models for normal and malignant breast tissue dielectric properties at microwave frequencies, IEEE Microw. Wireless Comp. Lett. 17 (12) (2007) 822–824.

[34] V. Y. Sohn, Z. M. Arthurs, J. A. Sebesta, T. A. Brown, Primary tumor location impacts breast cancer survival, Am. J. Surg. 195 (5) (2008) 641–644.

[35] E. I. Blumgart, R. F. Uren, P. M. F. Nielsen, M. P. Nash, H. M. Reynolds, Predicting lymphatic drainage patterns and primary tumour location in patients with breast cancer, Breast Cancer Res. and Treatment 130 (2) (2011) 699–705.

[36] S. Rummel, M. T. Hueman, N. Costantino, C. D. Shriver, R. E. Ellsworth, Tumour location within the breast: Does tumour site have prognostic ability?, ecancermedicalscience 9.

[37] L. J. Esserman, et. al., Addressing overdiagnosis and overtreatment in cancer: a prescription for change, The Lancet Oncology 15 (6) (2014) e234 – e242.

[38] C. Scott, Performance measures for neyman-pearson classification, IEEE Trans. Inf. Theory 53 (2007) 2852–2863.

[39] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (3) (1995) 273–297.

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

[41] M. Davenport, The $2\nu$-SVM: A cost-sensitive extension of the $\nu$-SVM, Tech. Rep. TREE 0504, Dept. of Elec. and Comp. Engineering, Rice University, Houston, TX (Dec. 2005).

[42] R. Caruana, A. Munson, A. Niculescu-Mizil, Getting the most out of ensemble selection, in: Proc. Int. Conf. Data Mining (ICDM), 2006, pp. 828–833.