Statistical Analysis and Machine Learning Algorithms for RF Breast Cancer Screening

Collin A. Joseph

Electrical and Computer Engineering
McGill University
Montreal, Quebec, Canada

August 12, 2019

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Masters of Engineering

© Collin Joseph 2019

Abstract

The work of this thesis explores statistical and machine learning methods for anomaly detection in a novel low-power microwave breast cancer screening system. Reported dielectric contrast in the microwave frequency range between healthy and malign breast tissue is the main motivator behind the effort to design a time-domain radar-based prototype for safe breast screening. The microwave radar does not strive to yield a three-dimensional image of the breast interior. Instead, its aimed use would be for frequent monthly screenings which have the potential to detect a departure from the normal, hence increasing the chance of early detection and, in turn, successful treatment. The data used for the development of the algorithms was obtained either in controlled laboratory experiments on tissue-mimicking phantoms or in a clinical setting. Since the data is preliminary and scarce, the conclusions may be limited, but in the process of the algorithmic development, this work strives to take into account the nature of the signals and how they have been generated in this very new application. The following methods were adapted and applied to the data sets: simple statistical analysis to illustrate the differences in the data sets investigated in this work; discrete Fourier transform, short-time Fourier transform, empirical mode decomposition and ad hoc time domain analysis to derive effective feature extraction strategies for the radio-frequency radar scans; high-dimensional statistical hypothesis tests to investigate the characteristics of time-frequency features extracted; random search, random walk, simulated annealing, genetic algorithm and particle swarm derivative-free optimization algorithms to improve the computational efficiency of an ensemble cost-sensitive support vector machine classifier based on previous literature; and a forward step-wise ensemble selection algorithm to improve the predictive performance of the classifier. For each of the methods, the results were discussed in the light of the limitations of the collected data sets. Older data sets were found to have high signal amplitudes on average. Statistically significant differences between features extracted from scans with anomalies and scans without anomalies were only observed for scans of subjects with higher average permittivity. The time-frequency analysis features yielded superior predictive performance than feature extraction using dimensionality reduction by principal component analysis. The computational efficiency of the classifier was improved by a factor of at least 3.8 when optimization algorithms were used for hyperparameter selection, instead of an exhasutive grid search. With the data available, the forward step-wise selection algorithm did not improve the predictive performance as was anticipated.

Sommaire

Le travail de cette thèse explore des méthodes statistiques et d'apprentissage automatique pour la détection des anomalies dans un nouveau système de dépistage du cancer du sein par micro-ondes à basse puissance. Le contraste diélectrique rapporté dans la plage de fréquence des micro-ondes entre un tissu mammaire sain et malin est le principal facteur de motivation derrière l'effort de conception d'un prototype basé sur un radar à domaine temporel pour un dépistage du cancer du sein sans danger. Le radar à micro-ondes ne cherche pas à donner une image en trois dimensions de l'intérieur du sein. Au lieu de cela, son utilisation serait destinée à des dépistages mensuels fréquents susceptibles de détecter un écart par rapport à la normale, ce qui augmenterait les chances de détection précoce et, par conséquent, de traitement réussi. Les données utilisées pour le développement des algorithmes ont été obtenues soit dans des expériences de laboratoire contrôlées sur des fantômes imitant les tissus, soit dans un cadre clinique. Comme les données sont préliminaires et rares, les conclusions peuvent être limitées, mais dans le processus de développement algorithmique, ce travail s'efforce de prendre en compte la nature des signaux et la façon dont ils ont été générés dans cette toute nouvelle application. Les méthodes suivantes ont été adaptées et appliquées aux ensembles de données: analyse statistique simple pour illustrer les différences entre les ensembles de données étudiés dans le cadre de ce travail; transformée de Fourier discrète, transformée de Fourier à court terme, décomposition de mode empirique et analyse ad hoc du domaine temporel pour dériver des stratégies efficaces d'extraction de caractéristiques pour les balayages radar à fréquence radio; tests d'hypothèses statistiques de grande dimension pour étudier les caractéristiques des caractéristiques temps-fréquence extraites; recherche aléatoire, marche aléatoire, recuit simulé, algorithme génétique et algorithmes d'optimisation sans dérivées de particules, afin d'améliorer l'efficacité informatique d'un classifieur de machine à vecteurs de support sensible au coût basé sur la littérature précédente; et un algorithme de sélection d'ensemble pas à pas en avant pour améliorer les performances prédictives du classifieur. Pour chacune des méthodes, les résultats ont été discutés à la lumière des limites des ensembles de données collectés. Il a été constaté que les ensembles de données plus anciens avaient une amplitude de signal élevée en moyenne. Des différences statistiquement significatives entre les caractéristiques extraites des balayages avec anomalies et des balayages sans anomalies ont été observées uniquement pour les balayages des sujets ayant une permittivité moyenne plus élevée. Les fonctionnalités d'analyse temps-fréquence ont généré des performances prédictives supérieures à celles d'extraction de caractéristiques utilisant la réduction de dimensionnalité par analyse en composantes principales. L'efficacité informatique du classifieur a été améliorée d'un facteur d'au moins 3,8 lorsque des algorithmes d'optimisation ont été utilisés pour la sélection de l'hyperparamètre, au lieu d'une recherche par grille exhasutive. Avec les données disponibles, l'algorithme de sélection pas à pas en avant n'a pas amélioré les performances prédictives comme prévu.

Acknowledgments

I would like express my sincerest appreciation to Prof. Mark Coates and Prof. Milica Popović for their patience and steadfast guidance throughout my master's program. I am grateful for the opportunity to have worked with you both and benefited from your substantial expertise and wisdom.

I would also like to thank Lena Kranold (PhD Candidate) for her tireless work collecting data and the sisterly support and advice she shared to help me with challenges that I faced. Thanks to Dr. Yunpeng Li for his advice and helpful suggestions. Thanks to Müberra Özmen, Florence Robert-Regol, Jonas Scheller and Laure Abécassis for useful technical discussions and help with data collection. Thanks to Anne-Marie Zaccarin for her help translating the abstract.

Contents

List of Tables

List of Figures

List of Acronyms

ANOVA	Analysis of Variance
BIS	Best Individuals Selection
DFT	Discrete Fourier Transform
DNA	Deoxyribonucleic Acid
EMD	Empirical Mode Decomposition
FFT	Fast Fourier Transform
FSS	Forward Stepwise Selection
GA	Genetic Algorithm
GS	Grid Search
IIR	Infinite Impulse Response
IMF	Intrinsic Mode Frequency
MAV	Mean Absolute Value
MRI	Magnetic Resonance Imaging
PS	Particle Swarm
RF	Radio Frequency
RS	Random Search
RW	Random Walk
RX	Receiver
SNR	Signal-to-Noise Ratio
STD	Standard Deviation
STFT	Short Time Fourier Transform
SVM	Support Vector Machine
TDF	Time-Domain Features
TX	Transmitter
UWB	Ultra-wideband

Chapter 1

Introduction

1.1 Motivation

Breast cancer is one of the most common types of cancer in Canada. According to the Canadian Cancer Society, with the exception of skin cancers, it is the most common cancer among Canadian women and the second deadliest type of cancer to that demographic [1]. Early diagnosis is essential for effective treatment of the disease since the prognosis becomes less favorable the more the disease is allowed to develop and spread throughout the body [2].

Conventional breast cancer screening modalities lack either convenience, comfort or both for the patient being screened. The goal of the RF Breast Cancer Screening group at McGill is to provide a more comfortable and convenient supplement for these traditional methods. Specifically, the group hopes to develop a self-screening system using a radio-frequency radar system in conjunction with machine learning and digital signal processing algorithms. Figure 1.1 shows what the finished screening system might look like.

To this end, we want to gain a better understanding of the statistical properties of the radar scan data collected using this type of system. This is essential for the development of signal processing and machine learning detection algorithms for the system, as well as for guiding future iterations of the system's hardware if there is a need to accommodate for particular subject characteristics. Furthermore, this work investigates an efficient and robust detection algorithm, essential for a system such as this one which is intended to be used by non-experts in a non-clinical setting.

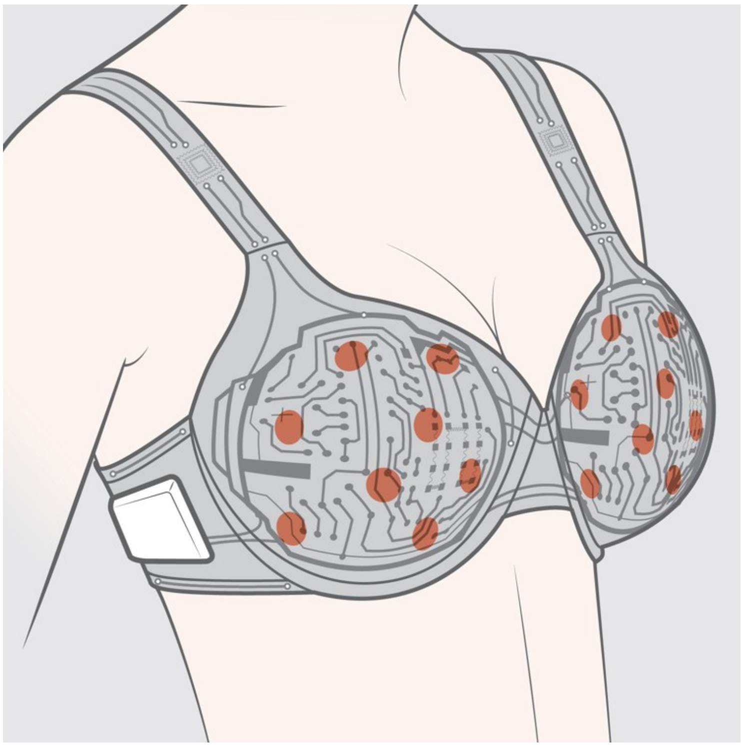


Figure 1.1: A concept illustration of what the final RF self-screening system might look like. The antennas would be housed in a comfortable bra. Screening would be performed with minimal discomfort to the patient. Image provided by Clinton Ford Illustrations.

1.2 Thesis Organization

This thesis is organized as follows:

- *Chapter 2 - Background and Literature Review*

  This chapter provides background knowledge necessary for understanding the work presented in later chapters as well as a review of recent literature in radio-frequency breast cancer monitoring. Firstly, an overview of conventional breast cancer screening modalities is presented. A review of recent work in radio-frequency screening research is then presented. Next, background information and literature related to the statistical hypothesis tests is given. Finally, the chapter offers background information about, and a review of the relevant machine learning and optimization methods used in the presented work.

- *Chapter 3 - Methods*

  This chapter provides a description of various data sets and technical methods used to obtain the results described in this thesis. Firstly, a description of the data sets analyzed in this thesis is given. Secondly, a description of the pre-processing methods used in this thesis is provided. This includes the signal windowing and band-pass filtering methods used. Thirdly a description of the various feature extraction methods used in this thesis is given. These methods include feature extraction algorithms using the short-time Fourier transform (STFT), discrete Fourier transform (DFT), empirical mode decomposition (EMD) and directly extracted time-domain features (TDF). Fourthly, a description of the high-dimensional statistical hypothesis tests used for data analysis is given. These tests include tests for variations in high-dimensional means and high-dimensional dispersions. Fifthly, a description of the machine learning classification algorithms used to perform the experiments described in this thesis are described. Sixthly, a description of various intelligent derivative-free search algorithms that may be applied to hyperparameter selection for machine learning problems are described. Seventhly, multiple ensemble pruning and selection algorithms for forming ensembles of machine learning models for classification problems are described. Finally, a description of the full breast cancer detection algorithms investigated in this thesis is given. This included descriptions of a previously published ensemble classifier and an explanation of various modifications made to the algorithm involving the aforementioned search algorithms and ensemble pruning and selection algorithms.

- *Chapter 4 - Results*

  This chapter presents the results of various experiments performed on radio-frequency breast monitoring data. Firstly, a numerical analysis of the peak absolute signal voltages of the data sets investigated in this thesis and a discussion of the analysis and its implications, is presented. Secondly, the results of statistical hypothesis tests performed on a recently collected clinical data set is presented. A discussion of these results is also presented. Thirdly, the results of classification experiments using new and old ensemble classifier algorithms are presented. A discussion of these results is also presented.

- *Chapter 5 - Conclusions and Future Work*

  This chapter presents the conclusions of the experiments conducted in this thesis and suggests potential future avenues of investigation.

1.3 Scope of Thesis and Contributions

This thesis presents the results of statistical hypothesis tests that investigate the variation of features extracted from the radio frequency radar scan signals of a recent clinical data set. Specifically, variations in the features of scans of subjects with and without anomalous tissue present and scan subjects of varying tissue density and combinations of these partitions are investigated to determine how easily these features can be used to distinguish the groups from one another and to gain insight into why or why not this is the case. The contributions of the author in this part of the thesis is the application of statistical tests to features extracted from the recently collected clinical data.

Additionally, the results of several machine learning experiments performed on a recent clinical data set as well as an older volunteer data set with artificially injected tumor responses. The first batch of experiments assess the benefit of using several hyperparameter search algorithms to identify good candidates for cost sensitive SVM ensemble selection, with regard to classification performance and classifier runtime. The second batch of experiments asses the benefit of using more complex ensemble selection strategies than simply selecting the best individuals, with regard to classification performance and classifier runtime. The contributions of the author in this section of the thesis is twofold. Firstly, proposing new ensemble classification algorithms using: intelligent hyperparameter search algorithms and ensemble pruning algorithms. Secondly, evaluating the performance of previously propose classifiers and the newly proposed classifiers on new and old phantom and clinical data sets.

Chapter 2

Background and Literature Review

2.1 Breast Anatomy and Tumors

Figure 2.1 from [3] shows a cross-section of a typical human female breast. The breast consists mostly of fatty (or adipose) tissue with fibrous (the ducts) and glandular tissue (the lobules) spread from the nipple throughout the breast.

Cancer is characterized by the uncontrolled growth and multiplication of abnormal cells that can accumulate and form tissues called tumors that prevent structures and organs within the body from functioning effectively. This behavior can be particularly catastrophic when it occurs in, or spreads to, vital organs because inhibition of their functionality then threatens the life of the affected organism [4], [5]. The majority of all breast cancers start in the ducts or the lobules of the breast [3]. It is therefore, essential to detect and treat tumors at their early stage prior to their spreading to other parts of the body.

2.2 Conventional Breast Cancer Screening

Several techniques for acquiring images of the tissue composition of breast are used for screening and diagnosis of breast cancer. They are reviewed here briefly, in the light of motivation for investigating additional diagnostic tools.

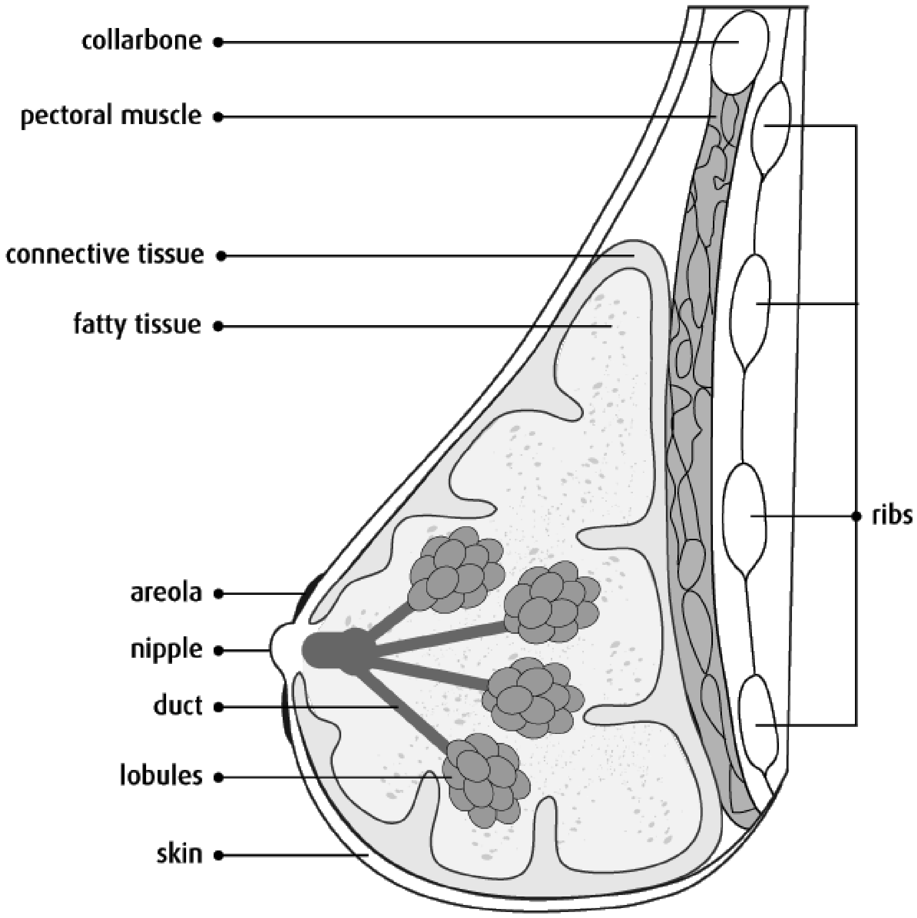


Figure 2.1: An illustration of a typical human female breast cross-section. The fibroglandular (the lobules and ducts) tissue, is typically spread throughout the adipose (fatty) tissue of the breast from the nipple. This image is modified from [3].

2.2.1 Mammography

Mammography is a common imaging technique for the screening and potential diagnosis of breast cancer [6]. It involves exposing a compressed breast to low-energy x-rays (photon energy in the range 10-20 keV) and observing the strength of the transmitted photons to construct an image of the breast tissues [7]. The x-ray photons are produced by an x-ray tube typically positioned above an image receptor, upon which the breast is held by the compression apparatus. The image receptor records the intensities of the photons transmitted through the breast tissue. The attenuation of the photon energy in each tissue within the breast is different [7]. Consequently, by measuring the spatial distribution of the attenuation of the photons, an image of the tissues within the breast can be constructed. The breasts are compressed to improve the quality of images acquired by the system. The compression reduces the transmission distance of the x-ray photons. This allows high resolution images to be captured using low-energy photons, minimizing the harmful ionizing radiation [7]. In addition, the compression of the breasts restricts movement, which reduces the amount of motion blur in the acquired image [7].

This screening technique has some disadvantages related to the level of expertise required to perform the screening and the health and comfort of the patient. Firstly, a mammogram requires a trained radiology technologist to operate the equipment, this increases the cost and reduces the accessibility of the procedure [6], [7]. Secondly, despite endeavors to reduce the photon energy as much as possible, repeated exposure to x-rays increases the risk of cancer development [6]. The photons used to generate the mammogram have the potential to damage DNA and cause carcinogenic mutations in otherwise healthy cells [8]. Thirdly, the breast compression causes discomfort to the patient.

2.2.2 Ultrasound Imaging

Ultrasound imaging is often used as a follow-up screening modality to mammography particularly if there is a need for additional screening or diagnosis of a specific area of the breast. Ultrasound imaging uses high-frequency sound waves (usually in the range 5-10 MHz) [9]. These are transmitted into the tissue under examination and the reflected sound waves are recorded. The characteristics of the reflected sound waves depend on the density of the tissue they were reflected by. This information can be used to construct an image [9]. Ultrasound imaging is a more comfortable experience for the patient since there is no need for pressure to be applied to the breasts. There is also almost no preparation required and no health risk due to radiation exposure [9]. However, this imaging technique also requires a real-time interpretation by a radiologist and suffers from non-specificity. Ultrasound results can be difficult to interpret for the purpose of routine screening [10].

2.2.3 Magnetic Resonance Imaging (MRI)

MRI is another imaging technique that may be used in the screening and diagnosis of breast cancer. MRI images are generated by differentiating tissues with varying concentrations of hydrogen atoms. When exposed to a high-level magnetic field, the hydrogen atoms are brought to a state where adequate radio-frequency exposure causes them to oscillate and emit energy recorded by the system [11]. MRI scans are very expensive procedures due to the cost of the equipment and also require special expertise and patient preparation to be conducted [12]. As a consequence, this imaging technique is less accessible than those previously discussed.

2.3 Low-Power Radio-Frequency and Microwave Breast Cancer Detection

Experimental evidence indicates that there is a substantial difference between the permittivities and conductivities of different types of breast tissue in the microwave frequency range, particularly between tumorous tissue and glandular or adipose tissue [13]–[16]. Microwave breast cancer detection and monitoring techniques take advantage of this contrast to extract information about the tissue composition.

When electromagnetic radiation in the microwave frequency range is incident on the boundary between two tissue types, the dielectric contrast causes a portion of the electromagnetic wave to be reflected. The larger the contrast between the two tissues that form the boundary, the larger the reflected portion of the incident radiation will be. By analyzing these scattered signals we can attempt to generate images of the breast tissue and determine whether there has been a change in the tissue content from one measurement to another.

2.3.1 Imaging-Based Approaches

A popular approach to breast cancer screening using microwave technology is the generation of images that are representative of the tissue profile of breasts to allow easy identification of suspicious tissue formations by human observers [17]–[19]. These images may also be analysed by software to further automate the screening process.

This imaging-based approach is the focus of several research groups. One of the most common approaches to imaging is known as confocal microwave imaging. This involves illuminating the subject (human breast or tissue phantom) with pulses from a radar antenna array, then synthetically focusing the recorded signals to estimate the amount of scattering that occurred at a finite grid of locations within the subject and thus create an image that is representative of tissue boundaries within the subject [17]. The University of Bristol group uses a multistatic radar system to generate images; they have conducted trials on simulated breasts, synthetic breast phantoms and human subjects [20]–[23]. The group at the University of Calgary uses a single antenna system to acquire signals for image generation. Experiments on synthetic phantoms and human subjects have been performed [24]–[28]. The group at the National University of Ireland Galway uses a multistatic system and confocal imaging techniques to generate images. Experiments on synthetic phantoms have been conducted [29]. In [30] and [31], Conceição et al. introduce a novel approach to imaging using machine learning classification algorithms (specifically, support vector machines) in conjunction with confocal microwave imaging techniques to generate images that show regions of tumorous tissue and regions of non-tumorous tissue.

Another approach to image generation is known as microwave tomography. This involves attempting to infer the dielectric profile of the subject from the scattering of the microwave signals that is observed. These approaches are typically done using frequency-domain systems. Researchers at Dartmouth College use microwave tomography to generate images, and have conducted experiments on phantom and clinical data sets [32], [33]. McMaster University's microwave imaging group recently has done work using microwave tomography to generate images of synthetic tissue phantoms [34]–[36]. The University of Manitoba's microwave imaging research team has also taken a tomography approach to imaging [37], [38]. Experiments have been conducted on synthetic breast phantoms.

2.3.2 Detection and Analysis using Machine Learning

An alternative approach is to focus exclusively on detection and analysis of anomalies revealed by the microwave screening systems rather than attempting to generate images [17].

Much resarch is focused on the differentiation between benign and malignant tumors based on their response to microwave illumination. In [39] three feature extraction methods are compared for the classification of tumors in simulated breast tissue models, using their radar target signatures. Principle component analysis and discrete wavelet transform features were found to outperform independent component analysis. In [40], a spiking neural network classifier is compared to linear discriminant analysis and found to yield superior performance for classification in dielectrically heterogeneous simulated breast models. Deep-learning techniques are applied to simulated breast models in [41] and found to perform well compared to other machine learning techniques. In [42], linear and quadratic discriminant analyses are applied to radar target signatures of synthetic tumor phantoms. In [43], Olivera et al. present a detailed investigation into several aspects of tumor classification using simulated breast models, including: antenna location, grouping of antenna signals at the classifier level, signal pre-processing and feature extraction strategies. In this work, random forest classifiers are used to perform the tumor classification.

Alternative analysis and detection endeavours using machine learning include localization and detection of abnormalities in the breast tissue. Sekkal et al. recently applied artificial neural network algorithms to tumor localization in simulated breast models [44]. Song et al. present an investigation of time-frequency feature extraction methods for detection of anomalies in augmented human breast scans from a multistatic radar system developed at McGill University in [45].

2.3.3 McGill University System

The McGill RF Breast Cancer Screening group uses a 16-element multistatic radar system which takes advantage of the aforementioned dielectric contrast between the tumorous and healthy breast tissues [46]. The system consists of an array of 16 antennas [47]–[49] arranged in a hemispherical frame as in [50] or in a prototype bra as shown in Figure 2.2. This array of antennas is used to perform radar scans of human breasts or synthetic breast phantoms. The radar scan is performed using a short-duration, ultra-wideband (UWB) pulse in the frequency range 2-4 GHz [51]. Figure 2.3 shows a schematic of the group's hardware system. The 16 antennas are activated in individual transmit-receive pairs to capture the radar signals, in other words, each antenna transmits a pulse while another antenna records the scattered and transmitted signals. This produces a total of 240 signals per scan, one signal for each uni-directional pair of antennas in the array.

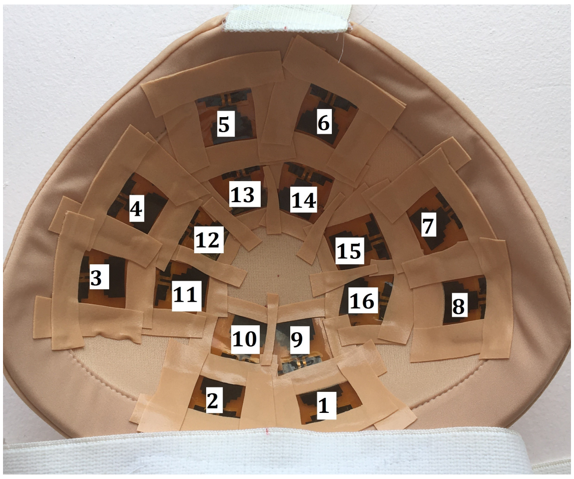


Figure 2.2: The McGill RF Breast Cancer Screening group's prototype bra using the ring antenna configuration. The flexible antennas are approximately arranged in concentric circles radiating outwards from the center of the bra. The 16 antennas in the radar array are labelled for detailed observation of the 240 signals resulting from all the transmitter-receiver combinations.

The McGill University research group is focused mainly on detection of departures from healthy baselines rather than the classification of the specific anomaly [46], [52], [53]. The group does not strive to create images of the tissues within the breast. Recent endeavours to detect such departures from healthy baselines using ensembles of cost-sensitive support vector machines have been quite successful [46]. Extensive work has been done to re-create realistic synthetic breast phantoms and use these phantoms to perform controlled experiments in a laboratory environment. A significant amount of clinical volunteer experiments have also been conducted to collect and analyse data from real human subjects [54].

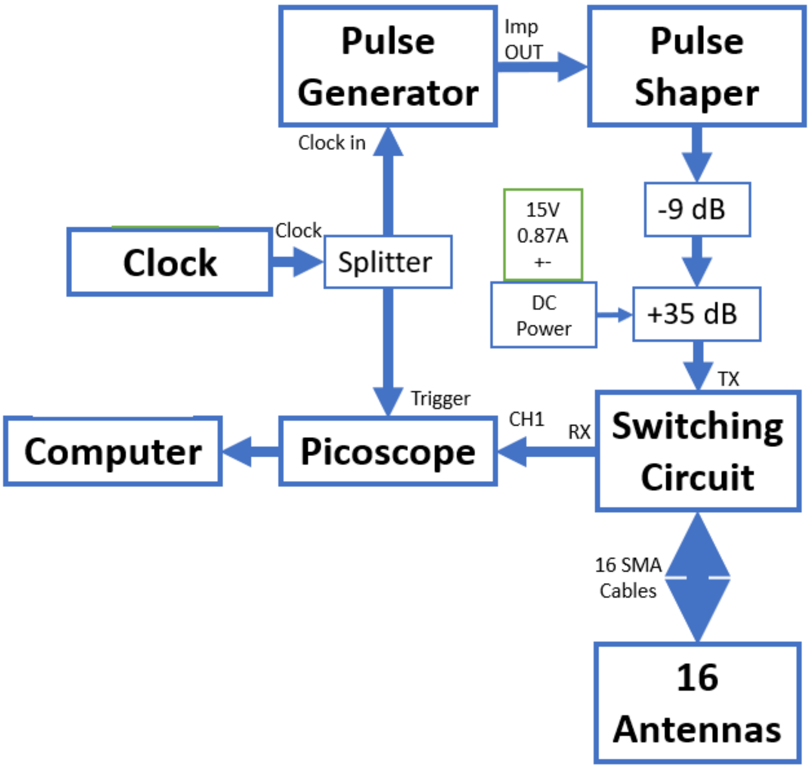


Figure 2.3: Block diagram of the McGill University time-domain radio-frequency radar breast monitoring system. The ultra wide-band pulse used in the radar system is generated by the pulse generator and pulse shaper, then passed through an attenuator and amplifier (for amplitude adjustment) to switching matrix that controls which pair of antennas in the 16 antenna array are transmitting and receiving. The received signal is recorded by the picoscope and stored on the computer.

Chapter 3

Methods

3.1 Data Sets

3.1.1 2014 Phantom Data Set

This data set [46] is a collection of scans of synthetic breast phantoms. These breast phantoms are designed to mimic the electrical properties of real human breast tissue [55]. There are 15 breast phantoms, 9 unique physical phantoms and 6 additional configurations that are achieved by rotating the original phantoms. 14 of the 15 phantoms allow a plug containing tumor tissue mimicking material or a plug mimicking adipose or glandular tissue to be inserted. This allows both tumor-less and tumor-bearing scans of these phantoms to be recorded. There are 292 scans in this data set; 10 scans of each possible configuration (tumor-less and tumor-bearing for each phantom) and two additional baseline (tumor-less) scans. In total, there are 140 tumor-bearing scans and 152 baseline scans. The scans were recorded at a sampling frequency of 200 GHz.

3.1.2 2014 Clinical Data Set

This data set is a collection of scans of healthy human breasts; the scans have selectively been injected with artificial tumor responses. The scans were originally collected from 12 volunteers over the span of several weeks as described in [46]. The injection of the artificial tumor response is achieved using approximate model of the microwave scattering in the breast tissue to create a signal response that represents the presence of a tumor in the breast tissue. The procedure is described in detail in [46]. There are 96 scans in this data set. Some of the 12 volunteers were able to participate in the study more frequently than others, consequently different numbers of scans were collected from each volunteer. These scans were recorded at a sampling frequency of 40 GHz.

3.1.3 2017 Clinical Data Set

This data set is a collection of scans of human breasts with and without anomalies present in the breast tissue, These scans were collected between October 2017 and April 2018 in a clinical setting at the Royal Victoria Hospital in Montreal, using the system described in Section 2.3.3. The data set includes scans from 39 individual volunteers for a total 71 scans. These scans were recorded at a sampling frequency of 160 GHz. At the time of recording the switching circuit of the system (see Figure 2.3) was highly susceptible to cross-talk between the transmitting and receiving lines and consequently produced a significant amount of noise before and during the UWB pulse in the recorded signals.

In addition to the recorded microwave scans, other data was recorded about each volunteer. The complete list is given in Appendix A. Additional notes related to the scanning procedure and the medical findings from the volunteer's consultation with a medical doctor are included. The recorded data most relevant for the purpose of this thesis were the volunteer's breast density and whether there was some anomaly in the volunteer's breast tissue. The density of each breast was labeled using the Breast Imaging Reporting and Data Systems (BI-RADS) mammographic density categories [56]. These categories describe the approximate percentage of fibroglanular tissue in the breast. The four categories are shown in Table 3.1.

Table 3.1: BI-RADS mammographic density categories. The numerical labels used to describe each breast density category are given.

Group Label	Fibroglandular Tissue Percentage
1	less than 25%
2	25-50%
3	51-75%
4	more than 75%

Table 3.2 shows the distribution of volunteer scans over breast density levels and

suspicion.

Table 3.2: Distribution of scans from clinical data set over breast density. The number of healthy (anomaly-free) and suspicious scans within each category is also shown. Due to ambiguous or insufficient labelling, 4 scans were excluded from the tallies shown in this table.

	Number of Scans		
Breast Density	Healthy	Suspicious	Total
1	4	3	7
2	10	9	19
3	13	15	28
4	7	6	13

3.1.4 2017 Phantom Data Set

This data set is a collection of scans of synthetic phantoms constructed in 2017. The data set was collected in February 2019 using the system described in Section 2.3.3. The data set is made with 508 scans in total, 208 baseline (tumorless) scans and 300 tumor-bearing scans. This data set was collected using an improved switching circuit that was not as susceptible to noise as the switching circuit used to collect the data in Section 3.1.3. A total of 9 different breast phantoms were used in the data set. For all scans, excluding 90 tumor bearing scans, phantoms were rotated 0, 45 and 90 degrees clockwise between successive scans to expand the measurement variation of the data set in a realistic way.

3.2 Signal Pre-processing

3.2.1 Signal Windowing

Due to the differences in the hardware used to collect the 2014 data sets and the more recent 2017 clinical data sets different windowing methods were necessary for each data set.

2014 Phantom Data Set

The raw signals were 4096 samples in length, however, only 2048 of these samples contained relevant signal information. Consequently, for these data sets, only the last 2048 samples of each scan signal were retained.

2014 Clinical Data Set

The scans in the version of this data set used in this thesis were already windowed to 1024 when the artificial tumor injection was performed. No additional windowing was used.

2017 Phantom Data Set

A fixed window was used for all signals. All scan signals started after 1500 samples and a window length of 1500 was found to be sufficient to consistently capture the entire recorded radar pulse.

2017 Clinical Data Set

Due to the noise introduced by the early switching circuit design, a more complex windowing procedure was necessary for the 2017 clinical data. The cross-talk on the switching circuit caused a large pulse to be recorded during signal transmission. This pulse, shown between sample indices 500 and 1500 in Figure 3.1 could easily be misinterpreted as the pulse that has been transmitted through the breast tissue and received by the receiving antenna. However, analysis of the expected recording system delay and hardware debugging indicated that this was not the case. A survey of the scan signals in this data set was conducted to manually obtain approximate windowing start samples that captured the appropriate portion of the signal. A fixed window size of 1500 samples was used for all scan signals. This window length value was empirically determined by a similar survey to sufficiently capture all the signal information that is thought to be useful. Figure 3.1 shows an example of this windowing method applied to a scan in this data set. The vertical dashed lines shown in red indicate the segment of the signal that is retained in the windowing procedure.

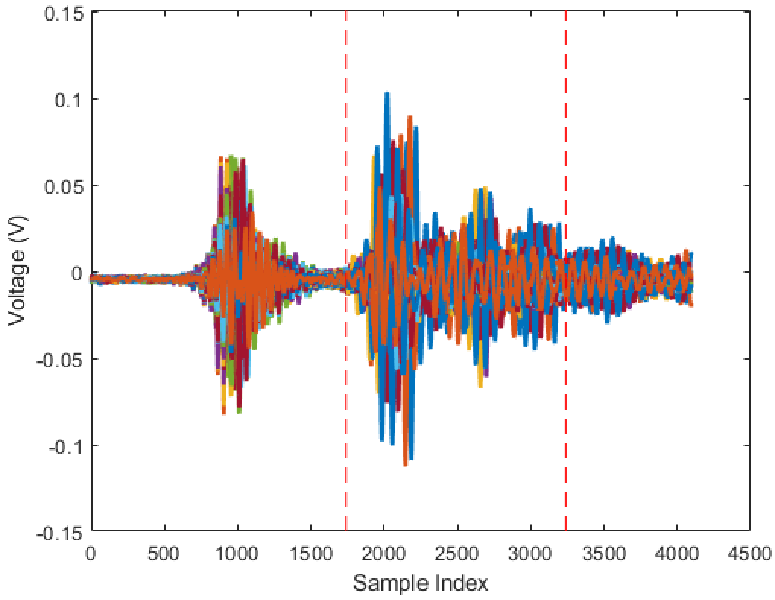


Figure 3.1: Example of windowing for 2017 clinical data. All the signals from a single scan are overlaid on a single axis. The first large pulse caused by the early switching circuit should not be included so the sample window should begin after it. Each scan was manually assigned an appropriate windowing start sample that was recorded and used for pre-processing.

3.2.2 Band-pass Filtering

In order to remove undesirable low and high frequency noise from the recorded signals a band-pass filter is applied to the 2017 clinical data. The 2014 data sets that were available for analysis were found to be usable without requiring any additional filtering. The scan signals are expected to carry most of their signal energy in the 2 to 4 GHz range since this is the frequency range of the UWB pulse used in the radar system. In practice it was found that a significant portion of the signal energy in the scans of the 2017 clinical data set was between 1.5 and 4 GHz. The exact reason for this spreading of the bandwidth is to be investigated in work outside of this thesis. Consequently a filtering pass-band of 1.5 to 4 GHz was selected to minimize the loss of useful information. Figure 3.2 shows the average magnitude spectrum of a scan from the 2017 clinical data set (after windowing as described in Section 3.2.1 has been performed). The vertical dashed lines shown in red mark the cut-off frequencies of the band-pass filter (1.5 GHz and 4 GHz). Most of the signal energy is within this frequency range. Figure 3.2 is cropped to show only frequencies upto 10 GHz for clarity. The filtering is done using a 6th-order Butterworth filter generated using the `scipy.signal` library [57].

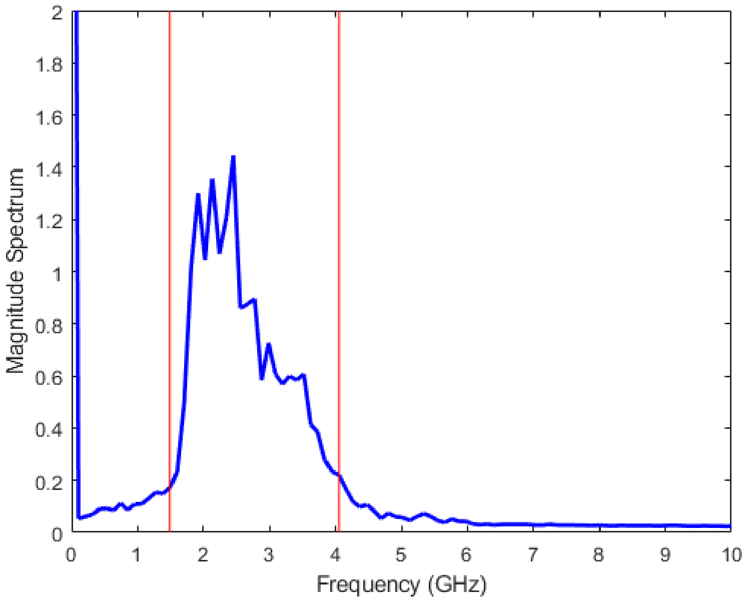


Figure 3.2: Average magnitude spectrum of a scan from the 2017 clinical data set. This was generated by computing the average magnitude spectrum over 240 signals comprising the scan. The vertical lines in red show the cut-off frequencies of the band-pass filter (1.5 GHz and 4 GHz). The low-frequency spike is due to noise and is not directly related to the electromagnetic wave transmission and scattering.

3.3 Feature Extraction

3.3.1 Principal Component Analysis

Principal component analysis (PCA) [58] is a dimensionality reduction method which works by identifying linear combinations of the original features that encapsulate as much information about the data set as possible. These liner combinations are referred to as the principal components of the data set. These components are identified by analysing the covariance of the original data set features. More specifically, these components are determined by solving a an eigenvector problem for the covariance matrix. The components are the eigenvectors of the covariance matrix and the eigen values are an indicator of how much information is encapsulated in each component. Only a subset of these components containing the majority of the data set information are retained. To generate features, data samples are projected onto the principal components to yield a “score”. In the context of this thesis, the covariance of the time domain signal samples is what is analyzed to yield principal components. For each signal, the 30 most significant principal component scores are retained.

3.3.2 Short-time Fourier Transform

The short-time Fourier transform (STFT) allows signals to be decomposed simultaneously in time and frequency and is useful for capturing variations in frequency over time. Time-frequency decomposition approaches to feature extraction for breast cancer detection have been attempted in [45] with limited success. Statistical features of the the short-time Fourier transform magnitude spectra of the scans were extracted and used for analysis. The STFT of each signal was computed using a Hanning window of length 256 samples with 32 samples of overlap between windows. The STFT was computed using the `scipy.signal` library [57]. Under this STFT configuration, the number of magnitude spectra produced for each signal (n_s) is represented by the following equation:

$$n_s = \lceil \frac{N_t}{256 - 32} \rceil \tag{3.1}$$

where N_t is the length of the time domain signal in samples. Consquently, there were $n_s = 10$ spectra for the 2014 phantom data set, $n_s = 5$ spectra for the 2014 clinical data set and $n_s = 7$ spectra for the 2017 clinical data set.

The statistical values computed for each STFT magnitude spectrum are described in Table 3.3. In the case of the STFT magnitude spectra, x represents the magnitude spectrum from which features are being extracted and N is the length of the spectrum. The statistics in Table 3.3 were computed using `scipy.stats` and `numpy` [57]. For the STFT features, there were 30 features per signal for the 2014 phantom data set which implies $30 \times 240 = 7200$ features per scan. There were 15 features per signal for the 2014 clinical data set resulting in $15 \times 240 = 3600$ features per scan. For the 2017 clinical data set, there were 21 features per signal which implies to $21 \times 240 = 5040$.

Table 3.3: Statistical features extracted from the time-frequency decompositions of the scan signals. The data sequence the statistics are extracted from is denoted x and its length is denoted N in the equations in this table.

Statistic	Equation
Mean Absolute Value	$\mu_{abs} = \frac{\sum_{i=1}^{N} \lvert x_i \rvert}{N}$
Standard Deviation	$\sigma = \frac{\sum_{i=1}^{N} \lvert x_i - \bar{x} \rvert}{N-1}$, $\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N}$
Kurtosis	$\kappa = \frac{\sum_{i=1}^{N} (\frac{x_i - \bar{x}}{\sigma})^4}{N}$

Since the statistics described in Table 3.3 are all different in nature it is unreasonable to expect that they would all be of comparable magnitude. Consequently, to avoid artificial skewing of feature importance because of relative magnitude, these features are scaled to a distribution with zero mean and unit variance by subtracting the mean (taken over all samples) and dividing by the standard deviation (taken over all samples).

3.3.3 Discrete Fourier Transform

Firstly, the discrete Fourier transform (DFT) of each signal was computed using a fast Fourier transform algorithm. This yields a frequency spectrum of length $N_t = 1024$, $N_t = 2048$ or $N_t = 1500$ samples, respectively. The variation in the number of samples was because of the different window lengths used for each data set. Since the time domain signal is real valued, only the first $\frac{N_t}{2} + 1$ samples are needed.

Secondly, the magnitude samples of the transformed signal in the frequency range 2-4 GHz are copied to a separate array. This is the frequency range of the transmitted microwave pulse and is consequently expected to contain useful information. The sample index range, $[i_{start},\ i_{end}]$, of the selected samples was computed using the following equations:

$$i_{end} = \lceil \frac{4 \text{ GHz} \times 2 \times (\frac{N_t}{2} + 1)}{f_s} \rceil \tag{3.2}$$

$$i_{start} = \lfloor \frac{2 \text{ GHz} \times 2 \times (\frac{N_t}{2} + 1)}{f_s} \rfloor \tag{3.3}$$

$$n_r = i_{end} - i_{start} \tag{3.4}$$

where f_s is the data set's sampling frequency. These magnitude spectrum samples were used as features to represent the signals from each scan. In this thesis, the phase information of the signals are ignored for simplicity. This method produces a different number of features per signal depending on the sampling frequency. For the 2014 phantom data $n_r = 22$ and consequently there were $n_r \times 240 = 5280$ features per scan. For the 2014 clinical data $n_r = 51$ and, with $n_r \times 240 = 12480$ features per scan. Finally, for the 2017 clinical data $n_r = 21$ resulting in $n_r \times 240 = 4800$ features per scan. Since all of these values (within each data set) were of the same nature, no scaling was necessary. The discrete Fourier transform was computed using the `numpy.fft` library [57].

3.3.4 Empirical Mode Decomposition

Empirical mode decomposition (EMD) is an adaptive time-frequency decomposition algorithm that decomposes signals into band-limited waveforms called intrinsic mode functions [59]. These features were extracted from each scan signal in a similar fashion to the STFT features in Section 3.3.2. Firstly, 4 levels of EMD were performed on each of the scan signals. This produced 4 IMFs. Secondly, the statistics described in Table 3.3 were extracted from each IMF. The features were also scaled to a distribution with zero mean and unit variance, as described in Section 3.3.2. This produced a total of 12 features per signal and $12 \times 240 = 2880$ features per scan. The EMD IMFs were computed using the `PyEMD` library [60].

3.3.5 Time-Domain Features

These features are statistical and numerical values extracted directly from the time-domain signals. A total of 4 features are extracted from each signal. The first two features are the mean absolute value and the standard deviation of the signal. These are described in Table 3.3. The third feature is the waveform length; this feature is taken from work by Hudgins et al. [61] in the area of myoelectric control and has been successfully used in electromyography pattern recognition tasks [62]. The waveform length (l_w) is defined by the following equation:

$$l_w = \sum_{i=1}^{N-1} |\Delta x_i| \tag{3.5}$$

where, x is the time-domain signal, N is the signal length and $\Delta x_i = x_{i+1} - x_i$. This feature is the sum of successive absolute sample differences and is essentially the cumulative waveform length. It encapsulates information about the waveform amplitude and frequency. The fourth feature is the number of extrema in the time-domain signal (denoted n_{ex}). This feature is computed using the algorithm in Algorithm 1. An extremum is identified as a point in the signal where the gradient changes direction. These last two features were chosen because the presence of a tumor in the breast tissue is expected to cause additional perturbation in the recorded signal compared to signals recorded in the absence of a tumor. Therefore, attempting to measure changes in the signal oscillation and complexity may be useful for detection. Since these features were expected to have significantly different value ranges, they were scaled to a distribution with zero mean and unit variance after they were computed, as in Section 3.3.2. This feature extraction method produced 4 features per signal which was equivalent to 960 features per scan.

Algorithm 1: Algorithm for computing the number of extrema (n_{ex}) of a time-domain signal. An extremum is identified as a point in the signal where the gradient changes direction.

```
Input: time domain signal $x$, signal length $N$
Output: number of extrema $n_{ex}$
Set $n_{ex} = 0$
Set $\Delta_{prev} = x[1] - x[0]$
for $i = 1$ to $(N - 1)$ do
    $\Delta_{new} = x[i + 1] - x[i]$
    if $sign(\Delta_{prev}) \neq sign(\Delta_{new})$ then
        $n_{ex} = n_{ex} + 1$
    end if
    $\Delta_{prev} = \Delta_{new}$
end for
return $n_{ex}$
```

3.4 Statistical Analysis Techniques

Comparing the statistical properties of features extracted from scans of subjects with different characteristics can aid in the development of detection algorithms and hardware by providing insight into how variations in subject characteristics affect the extracted features. The features extracted from the radio-frequency radar scans form multidimensional feature vectors. These feature vectors could be examined using tests for univariate data such as the T-test [63] or ANOVA [64] to analyze the features individually, but this would produce a large number of very specific results and would consequently be difficult to interpret since the hypothesis test results of individual features mean very little on their own. Instead, comparing the feature vectors in a more direct manner to obtain more concise results is more informative.

3.4.1 Hypothesis Tests for High-Dimensional Means

Statistical hypothesis tests such as the T-test or ANOVA work by comparing test statistics derived from the two populations to the probability distribution of the statistic when the null hypothesis is assumed to be true. We can determine the probability (the p-value) that the statistic we computed was drawn from the null distribution and thus determine the probability that the null hypothesis is true.

The T^2-test described by Harold Hotelling [65] is a generalization of the widely known T-test that is designed to compare multivariate data directly. However, this test is not very effective when the dimensionality of the individuals in each population (say, p) is much larger than the number of individuals in each population, (say, n) as is the case with our data. The T^2 test statistic is computed as follows:

$$T_H = (\bar{X}_1 - \bar{X}_2)^T S^{-1} (\bar{X}_1 - \bar{X}_2) \tag{3.6}$$

where, $\bar{X}_i$ is the mean vector of population $i \in \{1, 2\}$ and S is the pooled sample covariance matrix of the 2 populations. Because $p > n$, S is singular (non-invertible), therefore T_H cannot be computed. Proposed alternatives to the T^2 test involved using some alternative to T_H involving the sum of squared mean differences with additional scaling or biases such as those in [66], [67] and [68] to compensate for the relatively small population sizes. These sum-of-squares-based methods work best with data where the relatively large differences between the population means are present in a large number of the components of the mean vector (i.e., if the true mean difference vector comprises mostly non-zero values). In cases where the differences are present in only a few components of the mean vectors, a supremum test is more appropriate. A supremum tests considers only the largest component-wise difference between the population means when computing the test statistic. Such a test is described in [69].

Since we do not know what kind of differences to expect between the populations we are comparing, a more general approach is desirable. The adaptive sum of powers (`aSPU`) test described in [70] uses various sum of powers test statistics to evaluate the populations being compared. The general formulae for these statistics are shown in Equations 3.7 and 3.8.

$$L(\lambda) = \sum_{i=1}^{p} (\bar{X}_1^{(i)} - \bar{X}_2^{(i)})^{\lambda}, \text{ for } \lambda \in \mathbb{Z} \tag{3.7}$$

$$L(\lambda) = max_{1 \leq i \leq p} (\bar{X}_1^{(i)} - \bar{X}_2^{(i)})^2 / \sigma_{ii}, \text{ for } \lambda = \infty \tag{3.8}$$

where, $\bar{X}_1^{(i)}$ is the ith element of the mean vector of the first population and $\bar{X}_2^{(i)}$ is the ith element of the mean vector of the other population and σ_{ii} is an estimate of the pooled variance of the ith element. Equation 3.7 is similar to the statistic described in [66], but generalized for any value of λ. Equation 3.8 is based on the statistic described in [69]. For each value of λ the asymptotic distributions described in [70] are used to compute p-values for each statistic. The most powerful result (the smallest p-value) of all the tests executed is then selected to generate a final p-value for the adaptive test. The ordering of the test powers is inferred from the size of the p-value produced by each test. This adaptive test is implemented in the `highmean` R package and uses $\lambda = \{1, 2, 3, 4, 5, 6, \infty\}$ [70].

In the context of this thesis, each sample in the populations compared in this test was a concatenated vector of features from a subset of signals from a particular scan. The populations themselves were selected according to subject characteristics that were expected to affect the distribution of the extracted features (such as breast density and anomaly presence). The number of features and samples varied according to the experiment being performed. The null hypothesis of the test was that the means of both populations were the same.

3.4.2 Hypothesis Tests for High-Dimensional Dispersion

It can be advantageous to determine whether there is a statistically significant difference in the population dispersion (variance) of two mutually exclusive groups of data.

A method for comparing multivariate distributions is described by Marti Anderson in [71]. This method compares the average distance of the members of each group to their respective group centroids (means). Multiple distance metrics were proposed in [71] for the analysis of ecological data. Euclidian distance is used in the experiments described in this document for simplicity. The other distance metrics described in [71] were developed for specific biology applications and are unlikely to be advantageous in the context of this thesis. ANOVA is performed on the computed average distances to determine whether there is a statistically significant difference between how the two populations are spread around their means (their dispersion). The p-value of the generated tests statistics can be computed using a theoretical distribution, or by permuting the quantities that the test statistics are derived from a large number of times to generate a "permuted" null distribution. This test is implemented in Python as the `permdisp` test from the `QIIME` (quantitative insights into microbial ecology) software library. In each experiment, 999 permutations are used to generate a p-value from the permuted null distribution (the "permuted p-value") which is produced in addition to the theoretically determined p-value (the "observed p-value"). The populations compared using this test and the hypotheses are the same as described in Section 3.4.1.

3.5 Machine Learning Algorithms

3.5.1 Detection Algorithms

Support Vector Machines

Support vector machines (SVMs) [72] perform classification tasks by computing a hyperplane of maximal separation between classes. The hyperplane, of the form:

$\hat{f}(x) = x^T\beta + \beta_0$, is computed by solving an optimization problem which maximizes the margin between the hyperplane and each class. This may be formulated as a minimization problem of the form:

$$\begin{aligned}&\min_{\beta,\beta_0} \frac{1}{||\beta||} + C\sum_{i=1}^{n}\xi_i \\ &\text{subject to } \xi_i > 0, y_i(x_i^T\beta + \beta_0) \geq 1 - \xi_i, \forall_i\end{aligned} \tag{3.9}$$

where, x_i represents the training data samples, y_i represents the corresponding binary labels ξ_i represents how far each data sample is on the incorrect side of the margin, (as illustrated in Figure 12.1 of [72]), and C is a tuneable penalization parameter. Once the hyperplane of the form is computed new data samples are projected onto the plane and classified according to the sign of the projection (since this indicates which side of the hyperplane the sample is on). For problems where the classes are not linearly separable an additional function, called a kernel, may be applied to the data to project the data points to higher dimensions where linear separation may be easier, and a separating hyperplane may be generated more effectively. In this thesis we use a radial basis function kernel, which is parameterized by a variable γ as described in Section 12.3 of [72].

Cost-Sensitive Support Vector Machines

Control over how the SVM algorithm treats specific error types can be achieved by modifying the minimization problem described in Equation 3.9 to allow different penalty weightings for errors in each class, such as in [73]. The cost-sensitivity is achieved in this thesis by modifying Equation 3.9 as follows:

$$\min_{\beta,\beta_0} \frac{1}{||\beta||} + w_iC\sum_{i=1}^{n}\xi_i \tag{3.10}$$

where, the value of w_i is dependent on the true label of the ith data sample. Specifically, in this thesis, the positive class weight is fixed to $w_+ = 1$ while the negative class weight w_- is treated as a single tunable parameter. The implementation in the `scikit-learn SVC Python` class [74] is used in this thesis.

Ensemble Learning

It has been found that groups of models often exhibit better predictive performance than individuals [75], this ensemble approach has proven to be useful for RF breast cancer detection [45], [46], [52], [53]. The ensemble classification methods used in this thesis are described in more detail in Section 3.5.3 and Section 3.5.4.

3.5.2 Methods for Hyperparameter Selection

In [46], a grid search over more than 3000 hyperparameter configurations and up to 240 antenna pairs ($3000 \times 240 = 720,000$) is used to select the best cost-sensitive SVM models to be included in the final ensemble. While a grid search is the most thorough approach and is guaranteed to find the best models (global optima) for any criterion, it also requires massive computation.

Derivative-free optimization algorithms [76], [77] are useful for hyperparameter selection problems. They allow optimal or near-optimal solutions to be found without needing derivative information about the objective function that defines the optimization problem. This is ideal since the objective functions that guide hyperparameter optimization in machine learning problems are typically complicated and non-differentiable. In this thesis, the objective function used is an error metric, consequently, smaller values are better and the goal is minimization. The error metric used is described in more detail in Section 3.5.4. Unless stated otherwise, these algorithms operate on a discrete hyperparameter grid defined in Table 3.4.

Random Search (RS)

This method randomly selects a limited number of samples from the hyperparameter space and chooses the best of these. This allows superior runtime performance at the risk of not finding an optimum. This search method has been shown to be comparable and sometimes better than to more sophisticated search strategies [78]. In each iteration, a random hyperparameter configuration is generated by selecting values for each hyperparameter from a uniform distribution of values within the pre-determined range of each hyperparameter. The random configuration is then evaluated and then stored if the objective function score is sufficient for inclusion in the model library.

Random Walk (RW)

This method starts at a random point in the search space then chooses a neighbor to move to with a probability related to the fitness of the neighbors. The random walk implementation used in this thesis is based on the implementation described in [77]. The m neighbors are first sorted by objective scores in ascending order (the best neighbors first). The fitness of each point is related to its objective function score. The fitness f_i of the ith neighbor, given an objective score H_i; is defined as:

$$f_i = \frac{1}{iH_i} \tag{3.11}$$

these fitness values are used to perform a roulette wheel selection. Specifically, a cumulative probability q_i is assigned to each neighbor using the following equations:

$$q_i = \sum_{j=1}^{i} p_j \tag{3.12}$$

$$p_j = \frac{f_j}{\sum_{k=1}^{m} f_k} \tag{3.13}$$

a random number, r, is then drawn uniformly from the range [0,1]. If $q_{i-1} > r > q_i$ then neighbor i is selected as the next step in the random walk. The probability that a neighbor will be selected is therefore proportional to its fitness relative to the other neighbors. The algorithm terminates when the maximum number of objective function evaluations is exceeded.

Simulated Annealing (SA)

This method was originally proposed by Kirkpatrick in [79] and is still frequently used to solve NP-hard problems with non-differentiable objectives [80], [81]. The method combines a simple greedy algorithm with an exponentially decaying random exploration component. Starting at a random point in the search space (in this case the hyperparameter space), the algorithm attempts to move randomly through the space. Each randomly generated move is accepted with probability 1 if it leads to a better objective function score, or if the objective score is not better, with a non-zero probability determined by the current “temperature.” The probability of moving to a point with an inferior objective function score in the search space at step k of the search is defined as follows:

$$p_k = exp(\frac{-\Delta_{obj}}{t_k}) \tag{3.14}$$

where t_k is the current temperature and Δ_{obj} is the difference in objective function scores between the current point and its neighbor. The neighbor is accepted if a random number, r, drawn uniformly from the interval [0,1] satisfies $r \leq p_k$. The temperature parameter is intended to control the frequency of random exploration as the algorithm traverses the search space. A relatively large initial temperature is chosen to encourage random exploration at the beginning of the search. The temperature decays at each step of the algorithm according to some predefined function or "cooling schedule" to encourage local exploitation in later stages of the search. In this thesis, a geometric decay function defined by the following equation for a predetermined decay parameter $\alpha \in (0, 1)$:

$$t_{k+1} = t_k\alpha \tag{3.15}$$

the algorithm terminates when the maximum number of objective function evaluations is exceeded.

Genetic Algorithms (GA)

This method uses procedures inspired by evolution in nature to find optima in search spaces [82], [83]. Genetic algorithms are frequently used to solve optimization problems with non-differentiable objectives that cannot be solved exactly with polynomial time algorithms [84], [85]. The algorithm used in this thesis maintains a population of "chromosomes" representing potential solutions. In many implementations the chromosomes are binary representations of the problem solutions. In the implementation used in this thesis the chromosomes are simply arrays containing the hyperparameter configurations. Floating point chromosomes have been shown to work as well as or better than binary encoded chromosomes [86], [87]. In each iteration, a pair of chromosomes are selected with probability proportional to their relative fitness. This selection is performed using the roulette selection process described in the random walk description in this section. The selected "parents" are combined to produce offspring using the averaging method described in Chapter 6 of [86]. The parent chromosomes C_{p1} and C_{p2} form the offspring chromosomes C_{o1} and C_{o2} according to the following equations:

$$C_{o1} = \alpha C_{p1} + (1 - \alpha)C_{p2} \tag{3.16}$$

$$C_{o2} = \alpha C_{p2} + (1 - \alpha)C_{p1} \tag{3.17}$$

where α is a pre-determined mixing parameter. These "alleles" or components of each offspring are then mutated with a fixed probability and a dynamic parameter range determined by the number of iterations that have been performed. This approach is described in Chapter 5 of [86]. If an element v is selected for mutation, the mutation is performed according to the following equations:

$$v_{new} = \begin{cases} v - \Delta(t, v - LB) & \text{if } r_b = 0 \\ v + \Delta(t, UB - v) & \text{if } r_b = 1 \end{cases} \tag{3.18}$$

$$\Delta(t, y) = y(1 - r_v^{(1-\frac{t}{T})^b}) \tag{3.19}$$

where t is the current iteration count; UB and LB are the upper and lower bounds for the parameter, respectively; r_b is a randomly selected binary digit (which takes values 0 and 1 with probability 0.5); r_v is a randomly selected real number in the range [0,1]; b is a parameter to control the influence of the iteration count. These newly generated offspring are then compared to the existing population. If an offspring is better than the worst member of the population then the offspring is added to the population and the worst member is removed. The algorithm manipulates solutions using floating point operations, but at the end of each iteration the solutions are rounded to the nearest integer grid values. The search region is bounded by the upper and lower parameter limits of the discreet grid described in Table 3.4.

Particle Swarm Optimization (PS)

This method was originally proposed by Eberhart in [88]. It is inspired by the foraging behavior of birds, fish and other swarming species in nature. This optimization method is also frequently used to solve NP-hard problems with non-differentiable objectives [89], [90]. The algorithm tracks a population of "particles", each with its own position and velocity. Each particle also keeps track of the best solution it has found and the best solution that has been found by the entire population. At iteration, t, the location of the kth particle, X_k^t, is updated according to its current velocity, V_k^t. The objective function is then evaluated at the particle's new location and the individual and global best are updated if necessary. The particle's velocity is then updated according to the particle's current velocity, the particle's individual best solution (P_{bk}) and the global best solution (G_b). The velocity is updated according to the following equation:

$$V_k^{t+1} = wV_k^t + c_1 r_1 (P_{bk} - X_k^t) + c_2 r_2 (G_b - X_k^t) \tag{3.20}$$

where, t is the current iteration of the algorithm; c_1 and c_2 are parameters that control the individual and social influence on the particle's velocity; r_1 and r_2 are randomly generated real numbers in the range [0,1]; w is the particle's inertia. The algorithm manipulates solutions using floating point operations, but at the end of each iteration the solutions are rounded to the nearest integer grid values. The search region is bounded by the upper and lower parameter limits of the discreet grid described in Table 3.4.

3.5.3 Methods for Ensemble Selection

Ensembles of models often have more predictive power than individual models [75]. Ensembles of cost sensitive SVMs have been applied to radio frequency breast cancer detection with some success [46], [52], [53].

Forward Stepwise Selection

In [91], [92] Caruana et al. propose a forward stepwise selection (FSS) method for constructing predictive model ensembles. The ensembles are constructed by successively adding classifiers from a library of models that maximize the ensemble's performance on a validation or "hill-climbing" data set. They also propose several additional strategies to improve the performance of the algorithm.

Once a library of base models has been generated, the ensemble is constructed iteratively. First, it is initialized by selecting the best base model in the library, then on each successive step, each other model is added to the ensemble, evaluated on the validation set and then removed again. The ensemble that yields the largest improvement in the evaluation metric being used to guide the ensemble selection is then included in the ensemble permanently.

In addition to this basic procedure, Caruana et al. propose additional methods to improve performance [91]. The following additional strategies will be investigated in this thesis. Firstly, selection of base models with replacement is proposed. For an ensemble of fixed size this reduces the chance of base models reducing the performance of the ensemble. If replacement is not allowed, once the models that improve the ensemble's performance have been selected, there will only be models that degrade its performance available in future iterations. By using selection with replacement, the algorithm is able to avoid degrading the performance unnecessarily if there only a small number of useful models. Performing selection with replacement mitigates this situation since repeating good predictors is less likely to degrade the overall ensemble performance. Secondly, several of the best models in the library are used to initialize the ensemble instead of the single best model to reduce the chances of overfitting to the first model's prediction errors in the early stages of ensemble construction.

3.5.4 Cost-Sensitive Ensemble SVM

Original Ensemble Classifier

The original classification algorithm can be described as follows. Firstly, a subset of antenna pairs are selected based on their median peak amplitude over all scans in the data set. If median peak amplitude is below a predetermined threshold the antenna pair is not used for classification.

Secondly, the training data is separated into multiple train/validation folds. This separation is performed according to the subjects of the radar scans. Specifically, each validation set consists of scans from a single subject (i.e. a single human volunteer or a single synthetic phantom).

Thirdly, for each of the selected antenna pairs, a grid search of cost-sensitive

SVM hyperparameter values is then conducted using the average performance over the train/validation folds to evaluate each hyperparameter configuration. The hyperparameter configurations are evaluated using the Neyman-Pearson score metric [93]. This computed according to the following equation:

$$\hat{e}_{NP} = \frac{1}{\alpha}\max(0, P_f - \alpha) + P_m \tag{3.21}$$

where, α is the false-positive target rate, P_f is the false-positive rate and P_m is the false negative rate. The final evaluation score for each configuration is the average Neyman-Pearson score over all train/validation folds.

Once this is done for all selected antenna pairs, an ensemble of classifiers is selected from this library of cost-sensitive SVM models. The best 100 models over all antenna pairs and hyperparameter configurations are chosen (this selection procedure will hereafter be referred to as ”best individuals selection” or BIS).

Three hyperparameters are tuned for each individual model, the general error penalty (C), the kernel parameter (γ) and the negative class penalty weight (w_{-}). A total of 4620 hyperparameter configurations are considered. The range of values considered for each hyperparameter is shown in Table 3.4.

Table 3.4: Hyperparameter values considered when building the library of cost-sensitive SVM models for ensemble selection. There are 4620 possible hyperparameter combinations.

Hyperparameter	Value Range
C	$\{2^{-5}, 2^{-4} ... 2^{15}\}$
γ	$\{2^{-15}, 2^{-13} ... 2^{5}\}$
w_{-}	$\{1, 2 ... 20\}$

For prediction, an ensemble of “sibling” ensemble models is created by training an ensemble with identical hyperparameters for each train/validation split and training the models in the ensemble on the training samples of that split. The prediction of each ensemble is a majority vote over all the individual models in the ensemble and the final prediction of the classifier is a majority vote over all the ensemble models.

For evaluation of this classification algorithm and it’s proposed variations, the training data is divided into several train/test splits before hand and the predictive performance of the final classifier is evaluated using the these test sets. Additionally, by varying the false-positive target rate α a variety of false-positive and true-positive rates are observed for each classification algorithm. Using these values a receiver operating characteristic (ROC) curve can be generated for each algorithm.

Proposed Feature Extraction

While PCA [58] is the feature extraction method of choice in [46], a variety of other feature extraction methods will be used in this thesis. PCA and these proposed feature extraction methods are described in Section 3.3.

Proposed Hyperparameter Search

Instead of performing a grid search over all hyperparamerer configurations to build the model library from which the final ensemble parameters are selected, various hyperparameter searches are used to find a selection of relatively strong classifiers without having to exhaustively evaluate all possible hyperparameter configurations. The hyperparameter search methods are described in Section 3.5.2. In each hyperparameter search, the best configurations encountered are recorded and used to form the model library that the final ensemble is built from. The objective function used to guide the hyperparamter search is the average Neyman-Pearson score over all inner train/test folds as in the original classifier design. However, since each false-positive target requires its own search, the number of false-positive targets that will be investigated is limited. The target values will be $\alpha \in \{0.05, 0.1, 0.2, 0.5\}$. false-positive target rates above 0.5 (50%) are not useful in a practical setting.

Proposed Ensemble Selection

Instead of simply selecting the best classifiers from the model library to form the ensemble classifier, more complex selection algorithms can be used to build the ensemble. The selection algorithms that will be investigated are described in Section 3.5.3. The proposed method used the Neyman-Pearson score as the objective function used to guide selection. Consequently, only a limited number of false-positive targets will be investigated. Specifically, target values $\alpha = \{0.05, 0.1, 0.2, 0.5\}$. Due to the magnitude of the model libraries searched in the original implementation of the ensemble classifier (on the order of 500,000 models) and the necessity to process each false-positive target individually. it is extremely time consuming to perform selection using the proposed methods. Consequently, the model library will be limited to the best SVM configurations found for each antenna pair (at most 240 models) to make the experiments conducted more feasible.

Chapter 4

Results

4.1 Peak Absolute Voltage Analysis

The system diagram in Figure 2.3 indicates that in the current radar screening system, all amplification of signals occurs before the UWB pulse is transmitted by the antennas, previous iterations of the system were also configured in a similar way. Therefore, the strength of the transmitted, and consequently the recorded signals is determined largely by the amount of amplification that is applied before transmission. In addition, since the only other expected effect on the signal strength was the attenuation that occurs as the pulse was propagated through the breast tissue, it is reasonable to assume that the strength of the recorded signals is closely related to the signal to noise ratios (SNRs) of the signals.

The higher the SNR, the less likely that components of the signal caused by scattering in the breast tissue will be corrupted beyond recognition by noise. Therefore, it is expected that, the stronger the signals in a data set are, the more likely the difference between scans of without tumors and scans with tumors can be observed. The following analysis results were computed using the data sets after the pre-processing described in Section 3.2.

4.1.1 2014 Phantom Data Set

Before analysis, the signals in this data set were windowed as described in Section 3.2.1. Figures 4.1 to 4.4 show various measures relevant for our discussion later in this chapter and are calculated from the signals obtained with tissue phantoms.

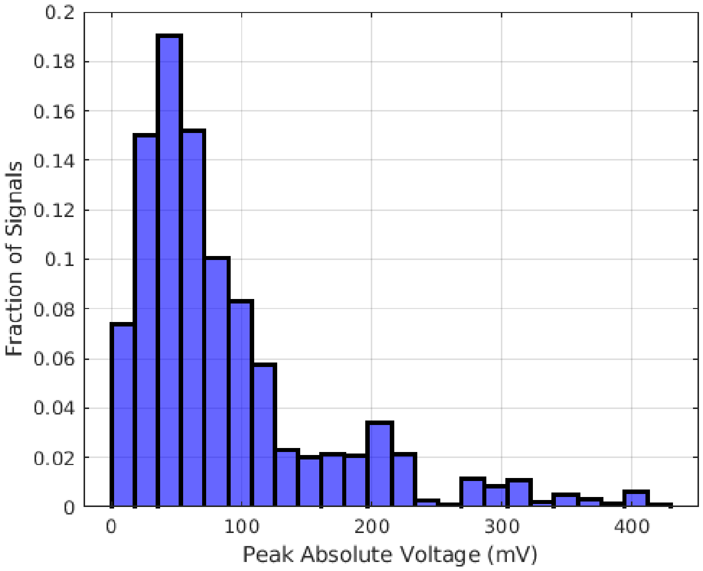


Figure 4.1: Histogram of peak absolute voltage values of signals in the 2014 phantom data set.

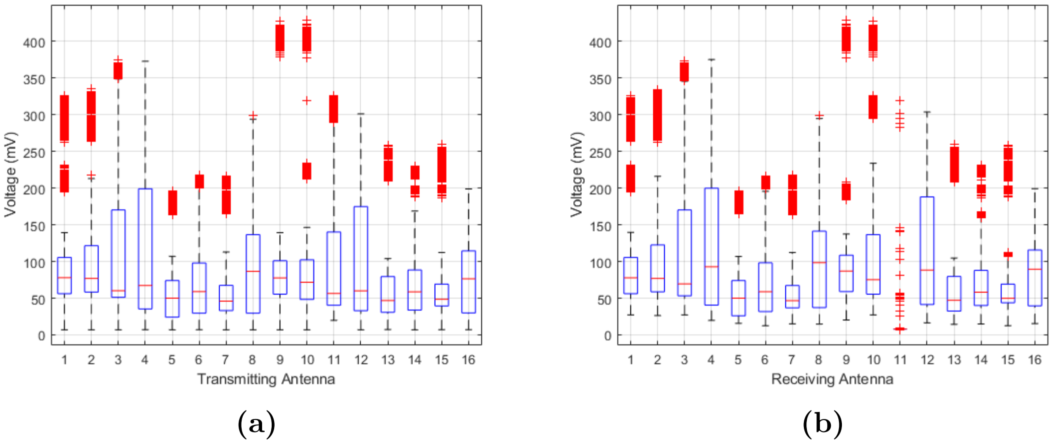


Figure 4.2: Box plots of peak absolute voltages of signals from the 2014 phantom data set grouped by (a) transmitting antennas and (b) receiving antennas. The red crosses indicate outliers from the distributions

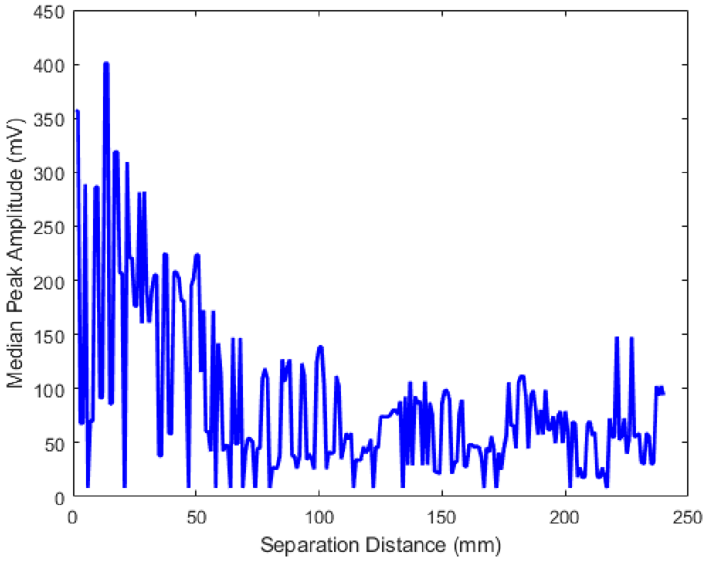


Figure 4.3: Median peak absolute voltage of antenna pairs in 2014 phantom data set sorted by antenna separation distance.

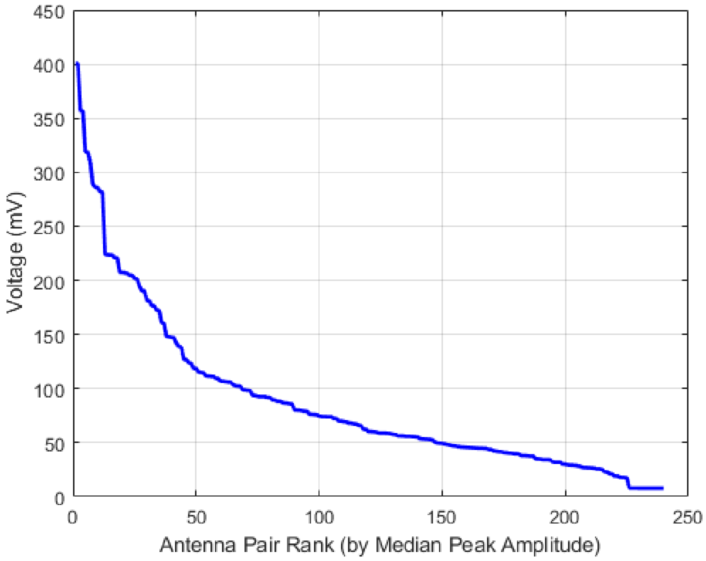


Figure 4.4: Sorted (in descending order) median peak absolute voltages of antenna pairs in 2014 phantom data set.

4.1.2 2014 Clinical Data Set

No additional pre-processing was performed on this data set before analysis. Figures 4.5 to 4.8 show measures needed for later analysis, calculated from the clinical data obtained in 2014.

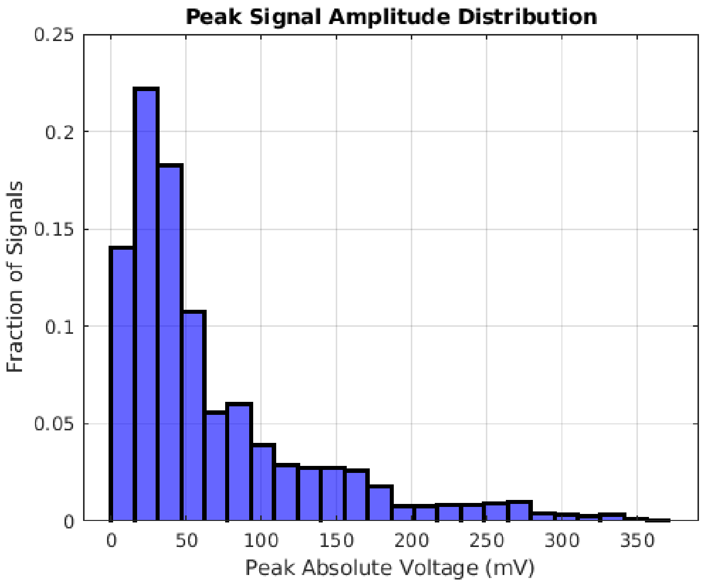
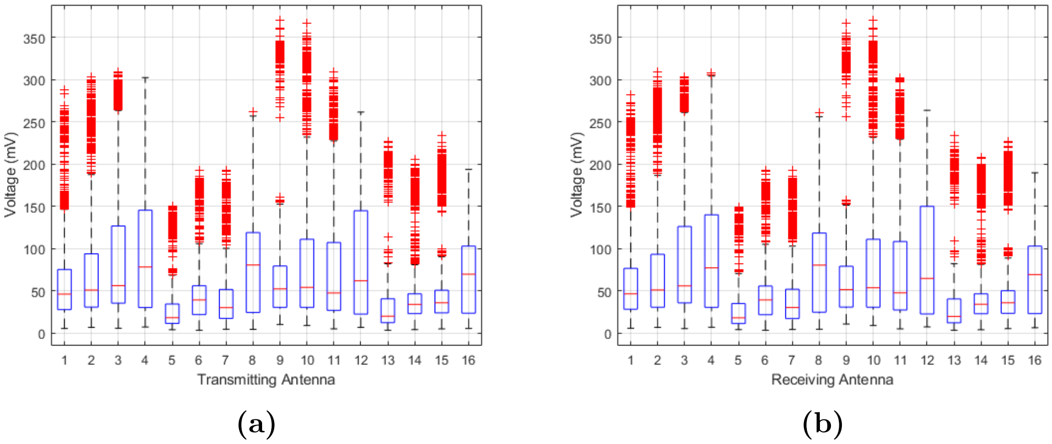


Figure 4.5: Histogram of peak absolute voltage values of signals in the 2014 clinical data set.



(a) **(b)**

Figure 4.6: Box plots of peak absolute voltages of signals from the 2014 clinical data set grouped by (a) transmitting antennas and (b) receiving antennas. The red crosses indicate outliers from the distributions

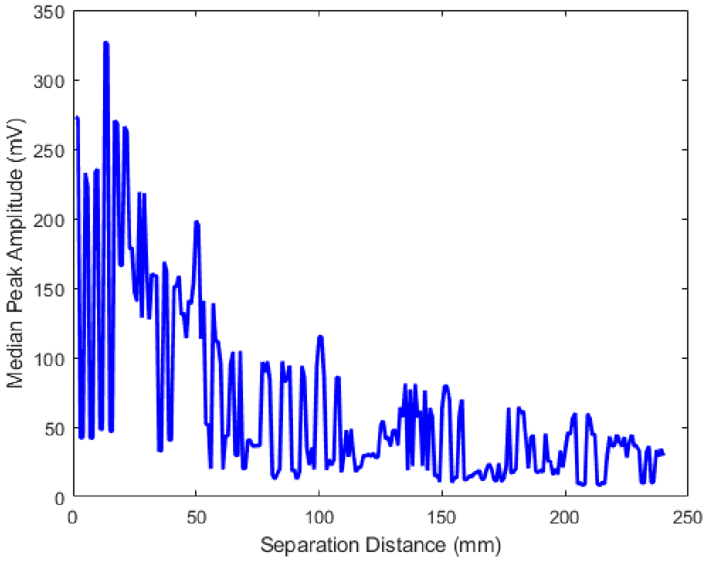


Figure 4.7: Median peak absolute voltage of antenna pairs in 2014 clinical data set sorted by antenna separation distance.

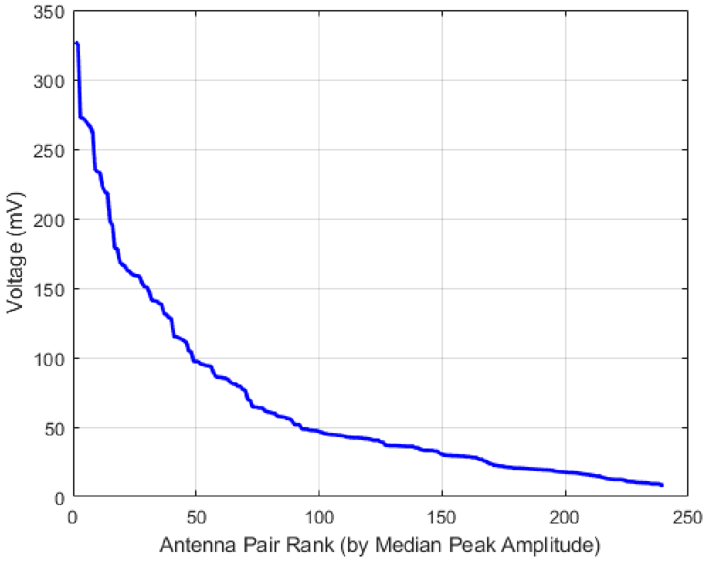


Figure 4.8: Sorted (in descending order) median peak absolute voltages of antenna pairs in 2014 clinical data set.

4.1.3 Discussion: 2014 Clinical and Phantom Data Sets

The peak amplitude analysis demonstrates a similar distribution of signal amplitudes between these two data sets. However there are some noticeable discrepancies. Figure 4.1 shows that the peak amplitude values are clustered around 50 mV in the 2014 phantom data set. However, Figure 4.5 shows that the signal amplitudes in the

2014 clinical data set the values are clustered around 25 mV. This variation may be a consequence of the subjects scanned in each data set. Synthetic breast phantoms may cause less attenuation of the signals than human breast tissue.

Both Figures 4.2 and 4.6 show significant variation in the peak signal amplitudes of each data set. The ranges of outlier values in Figure 4.2 is smaller than the ranges in Figure 4.6. This could be due to the difference in data set size. The 2014 clinical data set comprises only 96 scans while the 2014 phantom data set comprises 292 scans. Consequently, some of the outliers in Figure 4.6 may only appear to be outliers because of the small overall population size. Additionally, Figure 4.2 (b) shows that there is a flaw in receiving antenna 11. This antenna was damaged when the data set was collected and signals received by this antenna are omitted from most analyses in this thesis.

Figures 4.3 and 4.7 show very noisy decreasing trends in median peak amplitudes when they are sorted by distance. Both figures show a noticeable decreasing trend as separation distance increases, despite the noise. Figures 4.4 and 4.8 show steep drop-offs in median peak amplitude over the first 50 antenna pairs followed by a more gradual decay over the remaining 190 pairs in both data sets.

4.1.4 2017 Clinical Data Set

The signals in this data set were windowed, then filtered as described in Section 3.2. As in previous sub-sections, Figures 4.9 to 4.12 show measures for later analysis. They were calculated from the 2017 clinical data set.

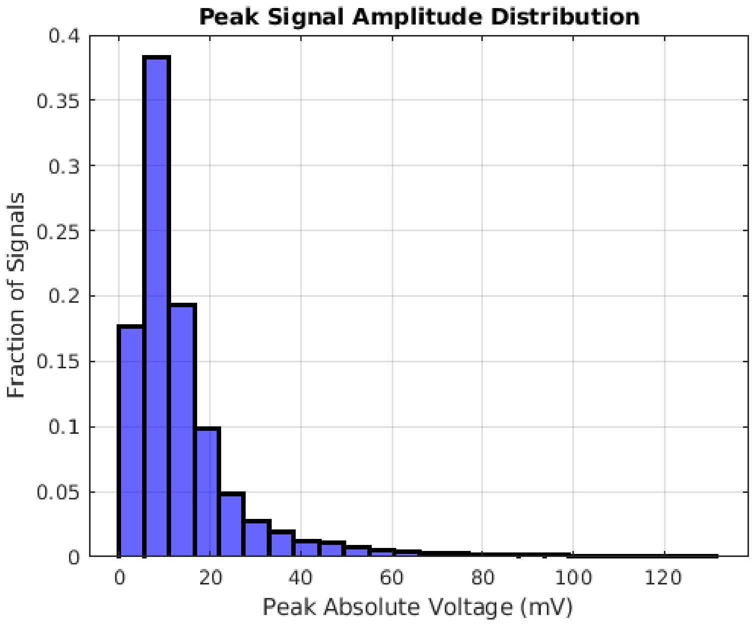
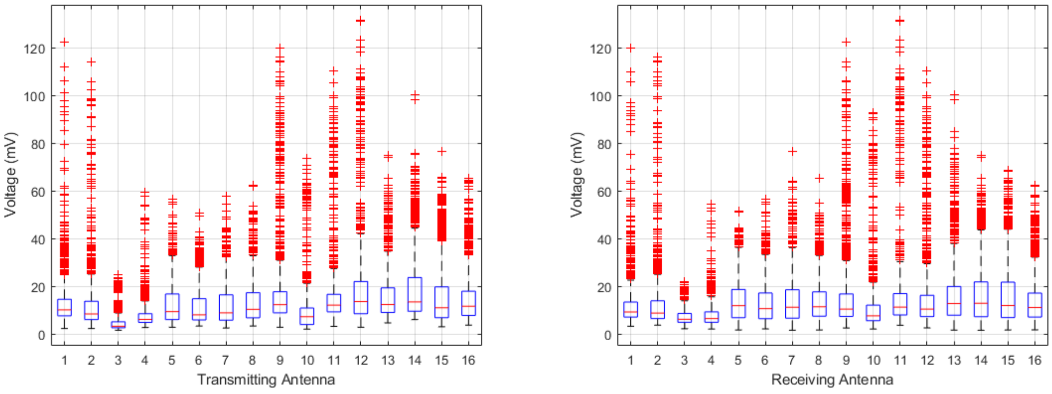


Figure 4.9: Histogram of peak absolute voltage values of signals in the 2017 clinical data set.



(a) **(b)**

Figure 4.10: Box plots of peak absolute voltages of signals from the 2017 clinical data set grouped by (a) transmitting antennas and (b) receiving antennas. The red crosses indicate outliers from the distributions

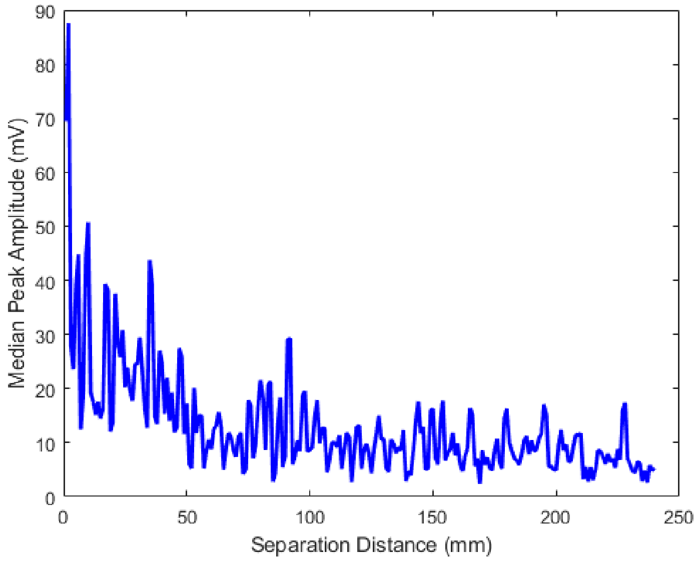


Figure 4.11: Median peak absolute voltage of antenna pairs in 2017 clinical data set sorted by antenna separation distance.

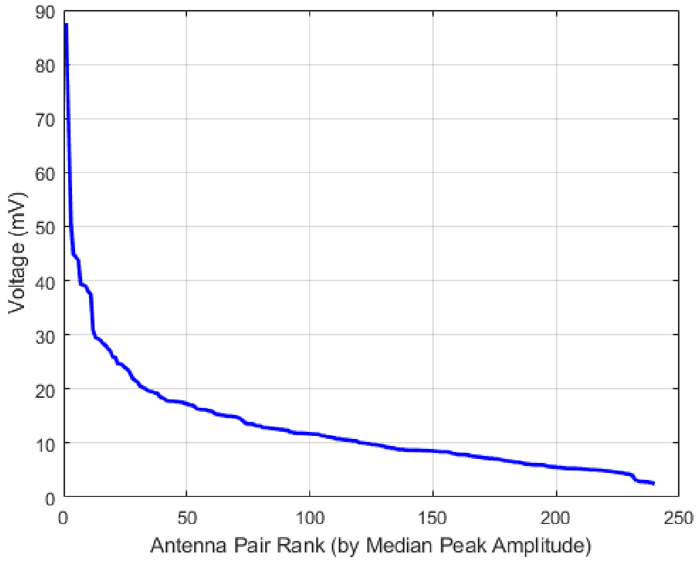


Figure 4.12: Sorted (in descending order) median peak absolute voltages of antenna pairs in 2017 clinical data set.

4.1.5 2017 Phantom Data Set

The signals in this data set were windowed and filtered as described in Section 3.2. As in previous sub-sections Figures 4.13 to 4.16 show measures for later analysis. They were calculated from the 2017 phantom data set.

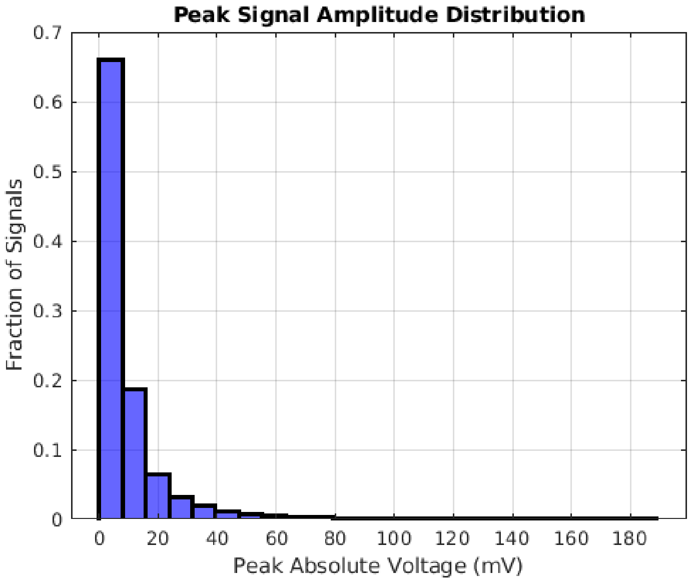
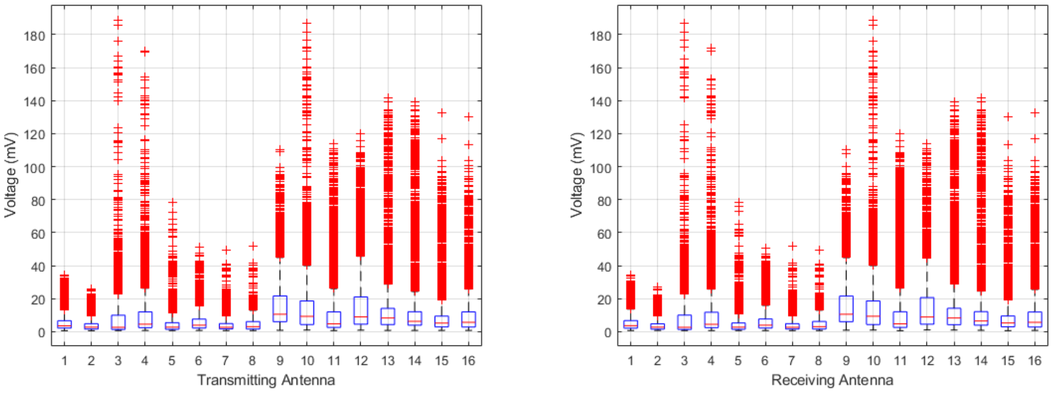


Figure 4.13: Histogram of peak absolute voltage values of signals in the 2017 phantom data set.



(a) **(b)**

Figure 4.14: Box plots of peak absolute voltages of signals from the 2017 phantom data set grouped by (a) transmitting antennas and (b) receiving antennas. The red crosses indicate outliers from the distributions

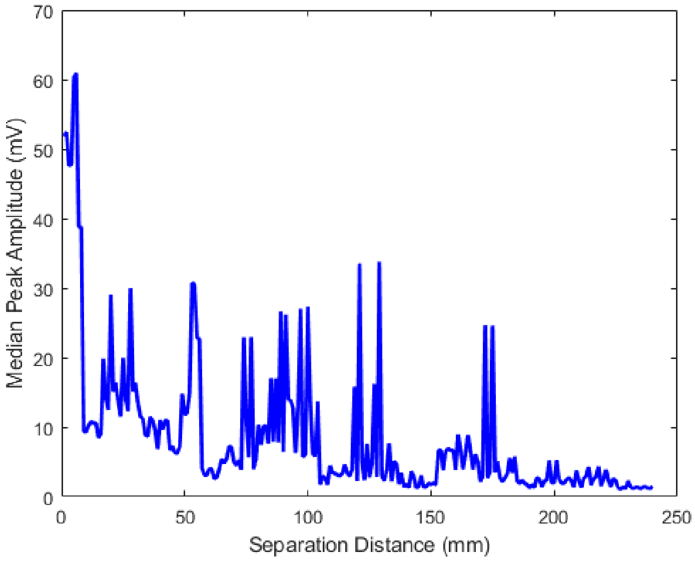


Figure 4.15: Median peak absolute voltage of antenna pairs in 2017 phantom data set sorted by antenna separation distance.

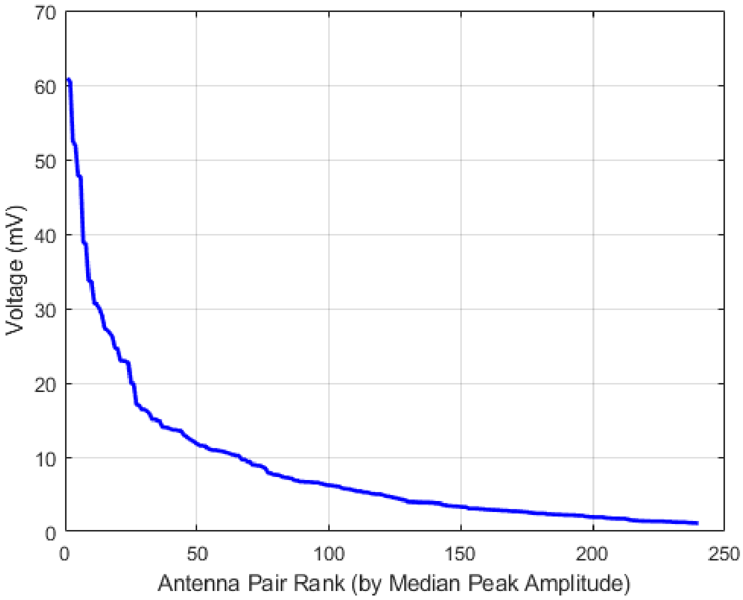


Figure 4.16: Sorted (in descending order) median peak absolute voltages of antenna pairs in 2017 phantom data set.

4.1.6 Discussion: 2017 Clinical and Phantom Data Sets

The peak absolute voltages of the signals in these data sets are significantly lower than the signals for the 2014 data sets. Figures 4.11 and 4.13 show that the majority of peak amplitude values are clustered around 5 or 10 mV and do not exceed 200 mV. On the other hand, the peak signal amplitudes of the 2014 data sets have histogram peaks around 50 and 25 mV and exceeded 350 mV. This suggests that the SNR of the 2014 data sets is higher than that of the 2017 data sets.

The discrepancies between Figure 4.10 (a) and Figure 4.10 (b) show that there was a lack of reciprocity between the transmitting and receiving lines of the switching circuit used to record the 2017 clinical data. In other words, some of the strongest transmitters are also some of the weakest receivers, e.g. antenna 14. Figure 4.14 shows a healthy reciprocity in the signal peak amplitudes once the new switching circuit is installed. The distributions of the peak amplitudes for the 16 antennas are similar when they are in both transmitting and receiving modes Sub figures (a) and (b) are almost identical suggesting that each pair of antennas performs similarly regardless of which antenna is transmitting and which one is receiving.

Figures 4.11 and 4.15 show a noisy decreasing trend with respect to increasing separation distance. However, the trend is noticeably less noisy than the 2014 data sets. In the 2017 data sets the median peak amplitudes drop off sharply from the maximum. In both the clinical and phantom data sets less than 50 antenna pairs have a median peak amplitude above 20 mV.

4.1.7 General Discussion

In all cases the peak amplitude analysis of these data sets show the similarity or difference in the hardware used to record the scans. In particular the peak amplitude distribution is closely related to the hardware. The 2014 data sets are very similar to each other, but very different from the 2017 data sets.

4.2 Statistical Analysis Experiment Results

It is useful to investigate how the data produced by the radio-frequency radar equipment responds to various types of scan subjects using the data sets described in Section 3.1. The difference between scans of healthy breasts and breasts with suspicious tissue present are of particular interest since this is closely related to the detection of anomalies in the breast tissue using machine learning algorithms. Other subject qualities that might affect the detection of anomalies, such as breast density, are also worthy of analysis. In this section, the properties of the 2017 clinical data set are investigated. Features extracted from the scan signals of this data set using methods described in Section 3.3 are analysed using high dimensional statistical hypothesis tests (described in Section 3.4) to gain insight into how these features are affected by the presence of anomalies in the breast tissue and variations in breast density. Similar tests are also performed on data from the 2014 phantom data set for comparison.

In these experiments we only consider the antenna pairs with the top 50 highest mean amplitudes. This is because in the RF radar system antenna pairs with stronger signals are expected to have better SNRs and would therefore be more likely to yield useful information when analysed. In addition, although the hypothesis tests described in Section 3.4 are designed to work effectively when the dimension of the population members (p) is higher than the size of the population (n), using extreme values of p that might be obtained if signals from the entire scan are concatenated, (see Section 3.3), with the relatively small $n = 71$ available is unlikely to yield more reliable results than using a subset of antenna pairs that are expected to be the most informative. Using the 50 strongest antenna pairs results in the dimension sizes shown in Table 4.1 for the feature sets described in Section 3.3.

Table 4.1: Length of feature vectors used in statistical hypothesis tests for 2017 clinical data set

Features	Dimension (p)
STFT	1050
DFT	1050
EMD	600
TDF	200

In the reported mean hypothesis test results, “SPU-λ” refers to the test for that particular value of λ in the adaptive sum of powers test and “aSPU” refers to the overall result of the test. In the reported dispersion tests, the “Permuted” test refers to the p-value obtained by random permutation and “Observed” refers to the p-value derived from the theoretical distribution. For all experiments we use a 5% confidence level for null hypothesis rejection.

4.2.1 Experiment 1: Healthy vs. Suspicious

In this experiment scans without anomalies (healthy) are compared to scans with anomalies (suspicious). Since fibroglandular tissue is known to have a higher permittivity than adipose tissue [14]–[16], the dielectric contrast between tumorous tissue and the rest of the breast tissue is likely to be lower for breasts with high concentration of fibroglandular tissue. This would decrease the amount of scattering observed in the radar scan. To account for this effect and to observe the degree to which it affects the difference between healthy and suspicious scan data, the scans are also partitioned into two breast density groups for this experiment. The first group comprises the two lower BI-RADS density groups (1 and 2), the other group comprises the higher density groups (3 and 4). The sizes of each population are shown in Table 3.2.

Experiment 1A: Healthy vs. Suspicious Means

In the first part of the experiment the means of the healthy and suspicious populations are compared using the adaptive sum of powers tests described in [70]. The null hypothesis for this experiment is that the means of each population are the same. The alternative hypothesis for this experiment is that the means of each population are different.

Table 4.2: Results (p-values) for healthy versus suspicious means hypothesis test for scans with breast density in groups 1 and 2 (low-density) from the 2017 clinical data set. P-values less than 0.05, which indicate that the result is statistically significant are shown in red.

Test	STFT	DFT	EMD	TDF
SPU-1	0.9	0.6	0.2	0.2
SPU-2	0.7	0.7	0.7	0.8
SPU-3	0.9	>0.9	0.7	0.9
SPU-4	0.7	0.7	0.7	0.7
SPU-5	>0.9	>0.9	>0.9	>0.9
SPU-6	0.6	0.6	0.6	0.6
SPU-∞	0.5	0.6	<0.001	0.2
aSPU	**0.9**	**>0.9**	**0.002**	**0.5**

Table 4.3: Results (p-values) for healthy versus suspicious means hypothesis test for scans with breast density in groups 3 and 4 (high-density) from the 2017 clinical data set. P-values less than 0.05, which indicate that the result is statistically significant, are shown in red.

Test	STFT	DFT	EMD	TDF
SPU-1	4e-05	1e-06	3e-10	5e-07
SPU-2	1e-16	4e-09	<1e-16	<1e-16
SPU-3	4e-16	1e-14	<1e-16	<1e-16
SPU-4	1e-16	7e-14	<1e-16	<1e-16
SPU-5	3e-14	<1e-16	<1e-16	<1e-16
SPU-6	3e-11	5e-16	<1e-16	<1e-16
SPU-∞	9e-07	4e-04	2e-04	1e-07
aSPU	**3e-16**	**<1e-16**	**<1e-16**	**<1e-16**

Table 4.4: Results (p-values) for healthy versus suspicious means hypothesis test for all scans in the 2017 clinical data set.

Test	STFT	DFT	EMD	TDF
SPU-1	0.6	0.08	0.1	0.4
SPU-2	0.8	0.7	0.7	0.8
SPU-3	0.7	0.7	0.6	0.8
SPU-4	0.7	0.7	0.7	0.7
SPU-5	0.9	>0.9	>0.9	>0.9
SPU-6	0.6	0.6	0.6	0.6
SPU-∞	0.8	>0.9	0.2	0.8
aSPU	**>0.9**	**0.4**	**0.4**	**>0.9**

Experiment 1B: Healthy vs. Suspicious Dispersions

In the second part of the experiment, the dispersions of each population were compared using the hypothesis test described in [71]. The null hypothesis for this experiment is that the dispersions of each population are the same. The alternative hypothesis for this experiment is that the dispersions of each population are different.

Table 4.5: Results (p-values) for healthy versus suspicious dispersion hypothesis test for scans with breast density in groups 1 and 2 (low-density) from the 2017 clinical data set.

Test	STFT	DFT	EMD	TDF
Permuted	0.7	0.8	0.4	0.8
Observed	0.7	0.8	0.4	0.8

Table 4.6: Results (p-values) for healthy versus suspicious dispersion hypothesis test for scans with breast density in groups 3 and 4 (high-density) from the 2017 clinical data set.

Test	STFT	DFT	EMD	TDF
Permuted	0.3	0.4	0.2	0.4
Observed	0.3	0.4	0.2	0.4

Table 4.7: Results (p-values) for healthy versus suspicious dispersion hypothesis test for scans all scans in the 2017 clinical data set.

Test	STFT	DFT	EMD	TDF
Permuted	0.2	0.5	0.07	0.3
Observed	0.2	0.5	0.07	0.3

Discussion

In Experiment 1A, populations containing only features extracted from healthy scans are compared to populations containing only features extracted from suspicious scans. Statistically significant results (results which indicate that the null hypothesis should be rejected) were only obtained for the tests involving low-density scans or high-density scans exclusively. For the tests involving low-density scans the only significant result was produced by the EMD features. Of the sum of powers tests performed, the most significant result was produced by the supremum test. The significance of the supremum test compared the lack of significance shown by the other tests suggests that a single feature is responsible for the statistical significance observed. The feature that produced the largest difference was the kurtosis of the 3rd IMF from the antenna pair TX14-RX8. For the tests involving only the high-density scans all the results were significant for all features with very high confidence in all cases. None of the results of the tests involving all of the scans were significant. This suggests that the variations in the population means for low and high-density scans are different, in fact they appear to be conflicting in all cases except that of the DFT features. In the majority of cases, the significance of the results for the combined data set is lower than the results for the isolated low and high-density scans. This variation in differences may arise because the difference in the breast tissue density changes the average permittivity of the breast tissue and consequently changes the way anomalies in the tissue cause scattering. The scattering caused by these anomalies may be captured more effectively by different features in low and high breast density scans. However, since a lower breast density would increase the dielectric contrast between anomalies such as tumors, it is expected that there would be more scattering present in these scans and thus larger differences would be observed in the features from the healthy and suspicious populations. In this case, the opposite appears to be true. Figures 4.17 to 4.24 show the distribution absolute mean features difference values between healthy and suspicious scan features, for each data set partition in the form of line graphs over feature indices and histograms. They indicate that, for all feature types, the difference between healthy and suspicious high-density breast scans is greater on average than this difference for low-density breasts. This supports the results observed in Tables 4.2 and 4.3.

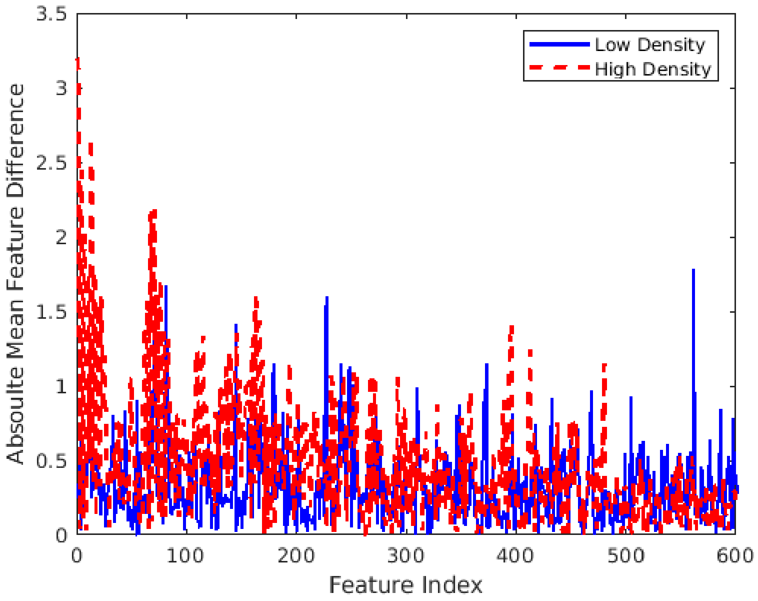


Figure 4.17: Absolute mean feature differences between healthy and suspicious scan STFT features extracted from the low-density scans only (in solid blue) and the high-density scans only (in dashed red).

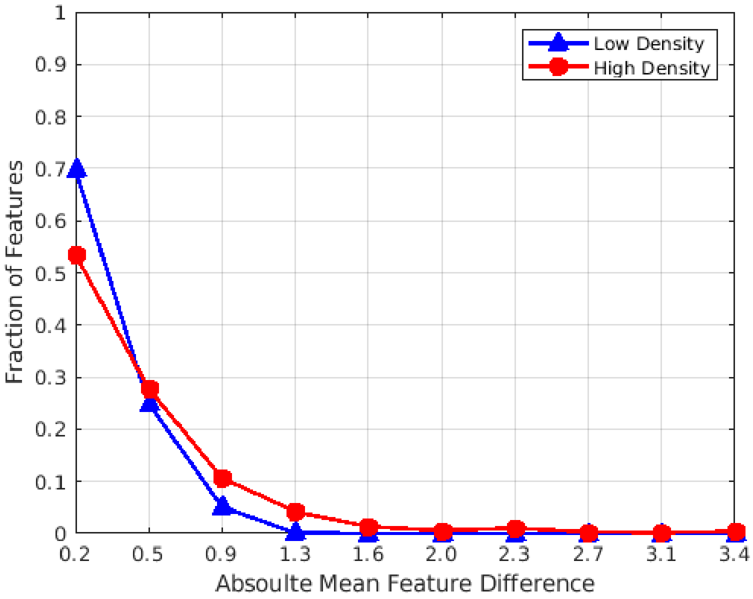


Figure 4.18: Histogram of absolute mean feature differences between healthy and suspicious scan STFT features extracted from low-density scans only (blue triangles) and high-density scans only (red circles). Values on the x-axis are the histogram bin centers.

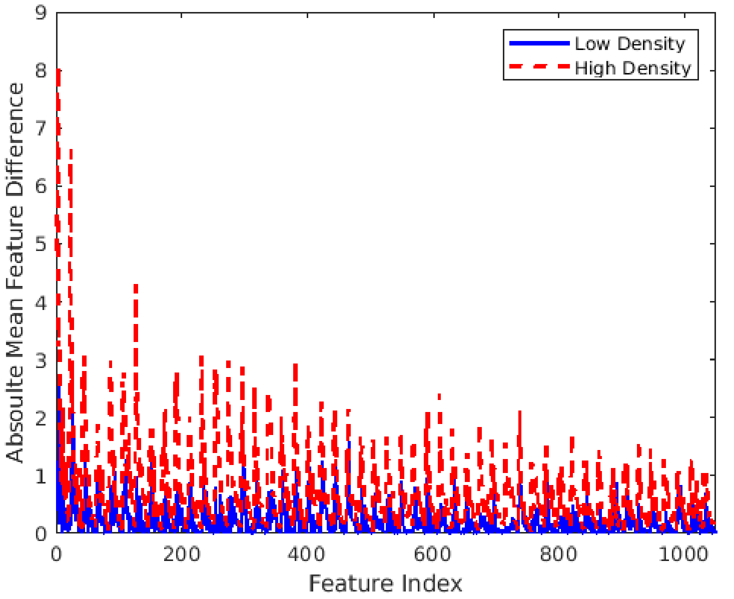


Figure 4.19: Absolute mean feature differences between healthy and suspicious scan DFT features extracted from the low-density scans only (in solid blue) and the high-density scans only (in dashed red).

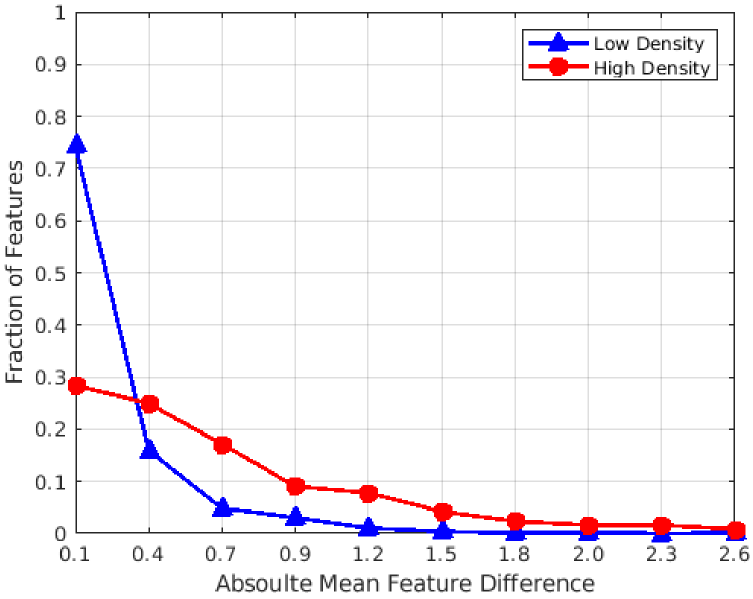


Figure 4.20: Histogram of absolute mean feature differences between healthy and suspicious scan DFT features extracted from low-density scans only (blue triangles) and high-density scans only (red circles). Values on the x-axis are the histogram bin centers.

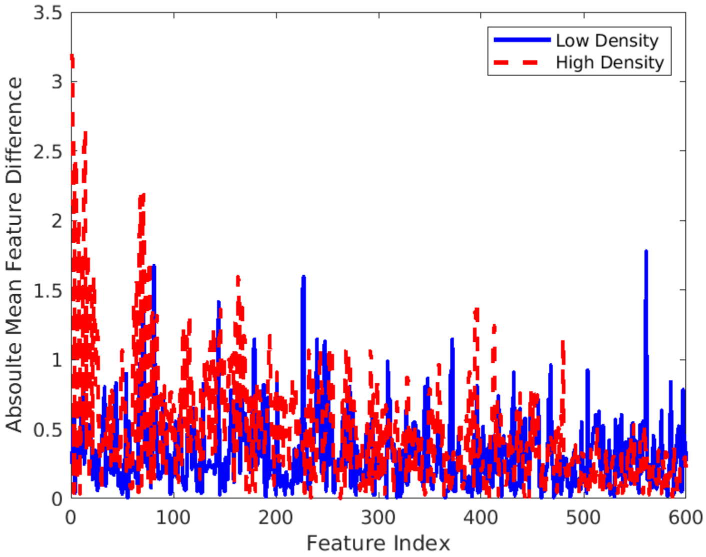


Figure 4.21: Absolute mean feature differences between healthy and suspicious scan EMD features extracted from the low-density scans only (in solid blue) and the high-density scans only (in dashed red).

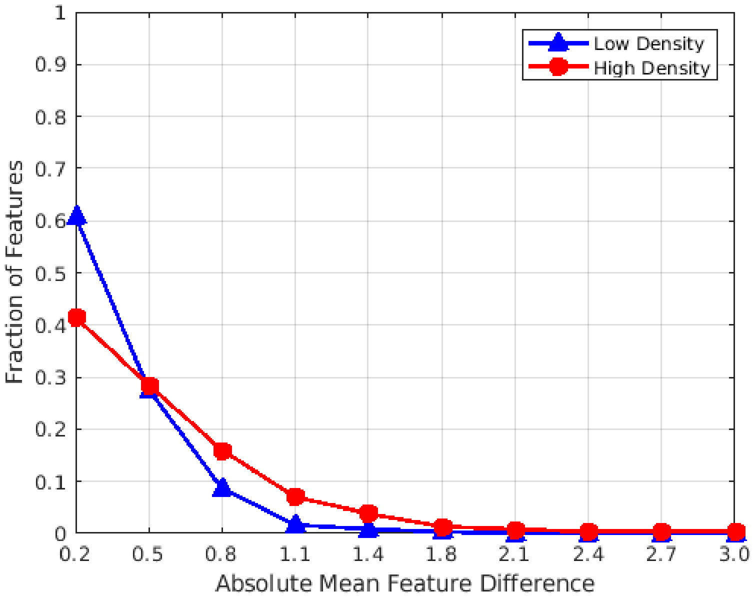


Figure 4.22: Histogram of absolute mean feature differences between healthy and suspicious scan EMD features extracted from low-density scans only (blue triangles) and high-density scans only (red circles). Values on the x-axis are the histogram bin centers.

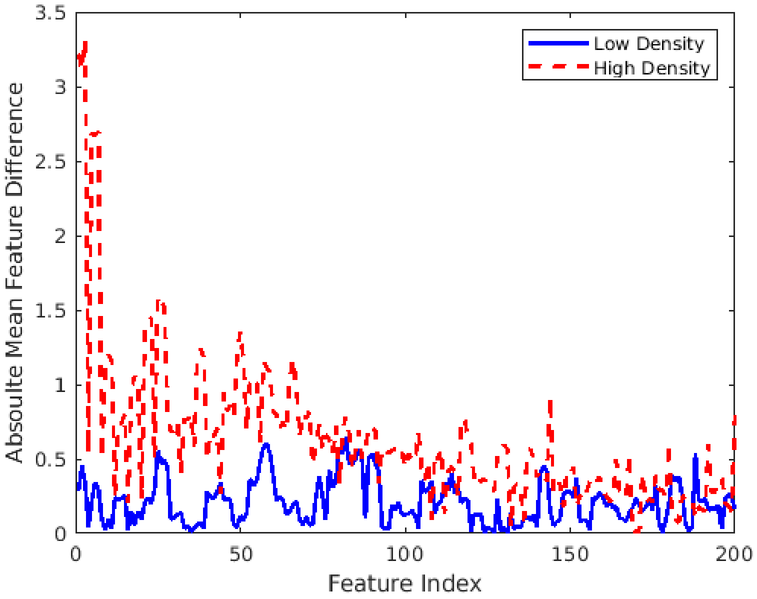


Figure 4.23: Absolute mean feature differences between healthy and suspicious scan TDF features extracted from the low-density scans only (in solid blue) and the high-density scans only (in dashed red).

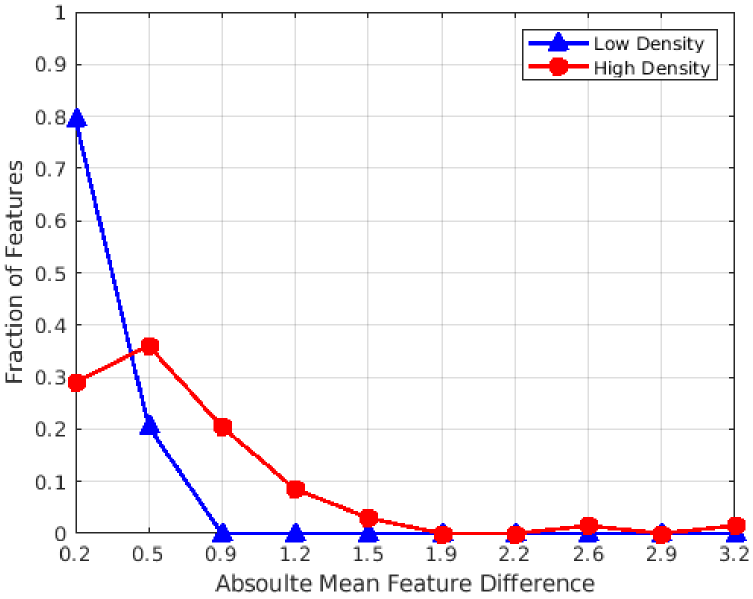


Figure 4.24: Histogram of absolute mean feature differences between healthy and suspicious scan TDF features extracted from low-density scans only (blue triangles) and high-density scans only (red circles). Values on the x-axis are the histogram bin centers.

In Experiment 1B, none of the results of the tests performed were statistically significant. The wide variation in the significance of the results for the low, high and combined density tests suggests as before that aspects of the feature dispersions that separate the healthy and suspicious scans are different for low and high-density scans.

This large difference between the results may be an indication that the multistatic radar system used to collect the scans is more suited for breasts of higher density and that a different approach must be taken for breasts with lower density. However, this conclusion should be drawn with reservation given the small populations in these tests (Table 3.2).

4.2.2 Experiment 2: Breast Density

In this experiment, features extracted from scans of low-density breasts are compared to features extracted from scans of high-density breasts. In each part of the experiment, to control for healthy versus unhealthy tissue, three tests were executed. Firstly, only the healthy scans were included in each population. Secondly, only suspicious scans were used. Thirdly, both healthy and suspicious scans were included in the populations.

Experiment 2A: Breast Density Means

In the first part of the experiment the means of the low-density (group 1 & 2) and high-density (group 3 & 4) populations are compared using the adaptive sum of powers tests described in [70]. The null hypothesis for this experiment is that the means of each population are the same. The alternative hypothesis for this experiment is that the means of each population are different.

Table 4.8: Results (p-values) for low-density versus high-density mean hypothesis test for healthy scans only from the 2017 clinical data set. P-values less than 0.05, which indicate that the result is statistically significant, are shown in red.

Test	STFT	DFT	EMD	TDF
SPU-1	0.3	0.8	0.5	0.8
SPU-2	0.5	0.3	0.4	0.4
SPU-3	0.9	0.4	0.7	0.8
SPU-4	0.6	0.4	0.6	0.6
SPU-5	>0.9	0.6	0.9	0.9
SPU-6	0.6	0.5	0.6	0.6
SPU-∞	0.04	0.03	0.01	0.2
aSPU	**0.2**	**0.08**	**0.03**	**0.5**

Table 4.9: Results (p-values) for low-density versus high-density mean hypothesis test for suspicious scans only from the 2017 clinical data set. P-values less than 0.05, which indicate that the result is statistically significant, are shown in red.

Test	STFT	DFT	EMD	TDF
SPU-1	4e-05	1e-03	1e-07	2e-05
SPU-2	4e-07	7e-08	2e-14	1e-12
SPU-3	2e-08	5e-05	2e-09	6e-08
SPU-4	6e-07	1e-02	1e-06	9e-06
SPU-5	9e-08	0.2	2e-03	2e-03
SPU-6	1e-05	0.3	3e-02	3e-02
SPU-∞	5e-10	5e-06	7e-07	3e-07
aSPU	**1e-09**	**4e-07**	**7e-14**	**6e-12**

Table 4.10: Results (p-values) for low-density versus high-density mean hypothesis test for all scans in the 2017 clinical data set. P-values less than 0.05, which indicate that the result is statistically significant, are shown in red.

Test	STFT	DFT	EMD	TDF
SPU-1	0.4	0.1	0.2	0.8
SPU-2	0.1	0.03	0.08	0.1
SPU-3	0.8	<0.001	0.2	0.6
SPU-4	0.4	0.001	0.3	0.4
SPU-5	>0.9	<0.001	0.6	0.7
SPU-6	0.5	<0.001	0.4	0.5
SPU-∞	0.03	0.02	0.4	0.1
aSPU	**0.1**	**<0.001**	**0.3**	**0.3**

Experiment 2B: Breast Density Dispersions

In the second part of the experiment, the dispersions of each population were compared using the hypothesis test described in [71]. The null hypothesis for this experiment is that the means of each population are the same. The alternative hypothesis for this experiment is that the means of each population are different.

Table 4.11: Results (p-values) for low-density versus high-density dispersion hypothesis test for healthy scans only from the 2017 clinical data set.

Test	STFT	DFT	EMD	TDF
Permuted	0.7	0.3	0.7	0.5
Observed	0.7	0.3	0.7	0.5

Table 4.12: Results (p-values) for low-density versus high-density dispersion hypothesis test for suspicious scans only from the 2017 clinical data set.

Test	STFT	DFT	EMD	TDF
Permuted	>0.9	0.06	0.8	0.8
Observed	>0.9	0.05	0.8	0.8

Table 4.13: Results (p-values) for low-density versus high-density dispersion hypothesis test for all scans in the 2017 clinical data set.

Test	STFT	DFT	EMD	TDF
Permuted	0.9	0.05	0.4	0.5
Observed	0.9	0.05	0.4	0.5

Discussion

In Experiment 2A, the only significant result for the tests involving only healthy scans was the result for the EMD features. This result was a consequence of a very significant supremum test result. The specific EMD feature that produced the maximum difference between the population means was the kurtosis of the 3rd IMF of the signal produced by antenna pair TX14-RX12. This result and the similar result obtained for the low-density scans in Experiment 1A seem to suggest that the kurtosis of the 3rd IMF is a particularly informative feature. For the tests involving only suspicious scans, all the results were significant. The reason for this is unclear considering the results of the other data set partitions. Table 3.2 shows that these population sizes are small, therefore it is relatively easy for this to be the result of random coincidence. For the tests involving all the scans in the data set, only the DFT features yielded a significant result. This may be because variations in breast density correspond to variations in the average dielectric properties of the breast tissue which may be readily observed in the frequency content of the recorded signals since the degree to which some frequencies are attenuated may vary significantly with the dielectric properties.

In Experiment 2B, the only significant result was produced by the DFT features. This, in addition to the results of Experiment 2A, particularly those in Table 4.10 suggest that DFT features may be the most appropriate features for differentiating between low-density and high-density breasts.

4.2.3 Experiment 3: Healthy vs. Suspicious for 2014 Phantoms

In this experiment, the tests in Experiment 1 are repeated on the 2014 phantom data set. Since this data set was collected in a more controlled environment than the 2017 clinical data and is known to be easy to classify using machine learning algorithms [46], [53], it is used to provide insight into these tests and what they may yield under more ideal measurement conditions. Table 4.14 shows the dimensionality of each feature set for the 2014 phantom data set.

Table 4.14: Dimensions of feature vectors used in statistical hypothesis tests for 2014 phantom data set

Features	Dimension (p)
STFT	1500
DFT	1100
EMD	600
TD	200

Experiment 3A: Healthy vs. Suspicious Means

In the first part of the experiment the means of the healthy and suspicious populations are compared using the adaptive sum of powers tests described in [70]. The null hypothesis for this experiment is that the means of each population are the same. The alternative hypothesis for this experiment is that the means of each population are different.

Table 4.15: Results (p-values) for healthy scan versus suspicious scan means in the 2014 phantom data set. P-values less than 0.05, which indicate that the result is statistically significant, are shown in red.

Test	STFT	DFT	EMD	TDF
SPU-1	0.1	0.004	1e-06	<1e-10
SPU-2	<1e-10	<1e-10	1e-04	<1e-10
SPU-3	<1e-10	<1e-10	0.5	<1e-10
SPU-4	<1e-10	<1e-10	0.4	<1e-10
SPU-5	<1e-10	<1e-10	0.8	<1e-10
SPU-6	<1e-10	<1e-10	0.5	<1e-10
SPU-∞	<1e-10	<1e-10	7e-10	<1e-10
aSPU	**<1e-10**	**<1e-10**	**2e-09**	**<1e-10**

Experiment 3B: Healthy vs. Suspicious Dispersion

In the second part of the experiment, the dispersions of each population were compared using the hypothesis test described in [71]. The null hypothesis for this experiment is that the dispersions of each population are the same. The alternative hypothesis for this experiment is that the dispersions of each population are different.

Table 4.16: Results (p-values) for healthy scan versus suspicious scan dispersions in the 2014 phantom data set. P-values less than 0.05, which indicate that the result is statistically significant, are shown in red.

Test	STFT	DFT	EMD	TDF
Permuted	0.001	0.001	0.9	0.001
Observed	1e-17	3e-34	0.9	3e-36

Discussion

In Experiment 3A, all feature sets yielded statistically significant results (from the `aSPU` adaptive tests). This suggests that the healthy and suspicious scan features are drawn from distributions with different means. In Experiment 3B, all the feature sets except the EMD features yielded statistically significant results. This suggests that the distributions of the healthy and suspicious scans are drawn from have different dispersions as well as means (indicated by Experiment 3A).

In both parts of Experiment 3, the EMD features yielded the least significant results. This could be due to the adaptive nature of the EMD time-frequency decomposition algorithm. Since the algorithm computes IMFs on a per-signal basis the information captured in each IMF might vary somewhat from signal to signal. Figure 4.25 shows the standard variation in the peak frequency of the first 4 IMFs extracted from the signals from selected antenna pairs of the 2014 phantom data set. A significant portion of the IMFs show large variation, specifically standard deviations between 10 and 30 GHz. In addition Figure 4.26 shows that half of the IMFs exhibit this high variance in peak frequency. This variation may act as noise that limits the difference that can be observed between tumor-bearing and tumor-free scans.

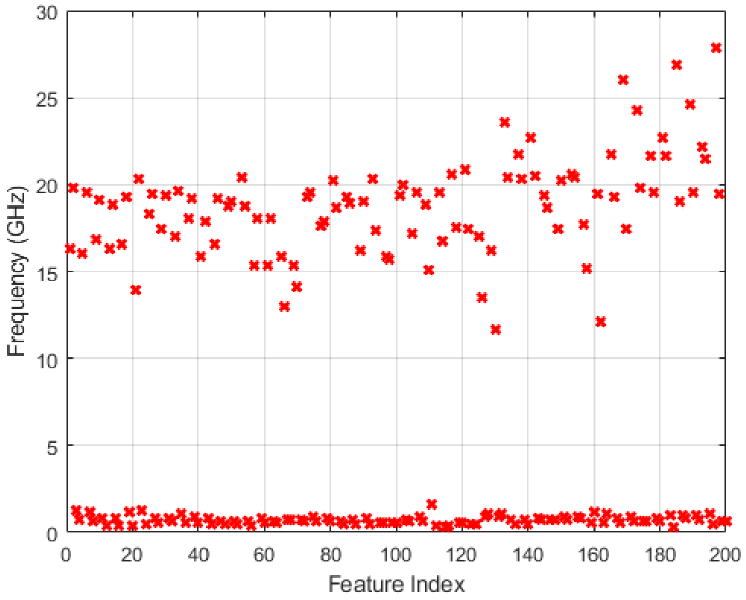


Figure 4.25: Standard deviation of peak frequencies of first 4 IMFs of signals from the selected antenna pairs in the 2014 phantom data set. Many of the IMFs demonstrate a large variation in peak frequency, in the range of 10 to 30 GHz.

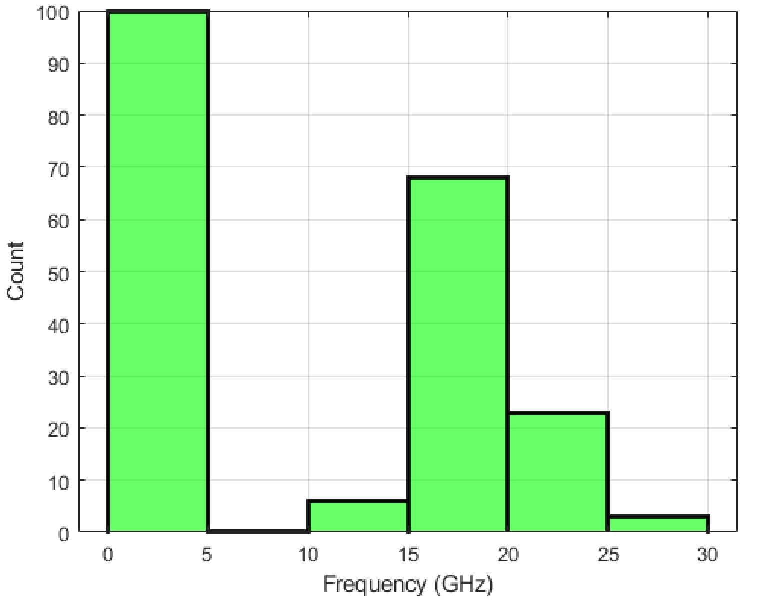


Figure 4.26: Histogram of standard deviation of peak frequencies of first 4 IMFs of signals from the selected antenna pairs in the 2014 phantom data set. Half of the IMFs have a peak frequency standard deviation of over 10 GHz.

Overall, these Experiment 3 results indicate that each of these feature extraction methods encapsulate information that can be used to differentiate between scans with anomalies present and scans without anomalies present. However, Experiment

1 shows that they were not as effective when applied to the 2017 clinical data set. This may be due to poor signal quality resulting from the relative lack of environmental control the 2017 clinical data set was collected under.

4.2.4 Experiment 4: Healthy vs. Suspicious for 2017 Phantoms

In this experiment, the tests in Experiment 1 are repeated on the 2017 phantom data set. As with the 2014 phantom data set, this data set was collected in a more controlled environment than the 2017 clinical data. However this data was collected with very similar hardware to the 2017 clinical data set with the exception of the switching circuit and the ring bra. These results are intended to explore how the results change in a controlled environment with a larger number of scans and synthetic breast phantoms instead of human breasts. Since the data windowing size and sampling frequency of this data set are the same as those of the 2017 clinical data set, the dimensionality of the feature vectors are the same as those in Table 4.1.

Experiment 4A: Healthy vs. Suspicious Means

In the first part of the experiment the means of the healthy and suspicious populations are compared using the adaptive sum of powers tests described in [70]. The null hypothesis for this experiment is that the means of each population are the same. The alternative hypothesis for this experiment is that the means of each population are different.

Table 4.17: Results (p-values) for healthy scan versus suspicious scan means in the 2017 phantom data set.

Test	STFT	DFT	EMD	TDF
SPU-1	0.8	>0.9	>0.9	>0.9
SPU-2	0.9	0.9	0.9	0.9
SPU-3	>0.9	>0.9	>0.9	>0.9
SPU-4	0.8	0.7	0.8	0.7
SPU-5	>0.9	>0.9	>0.9	>0.9
SPU-6	0.6	0.6	0.7	0.6
SPU-∞	>0.9	>0.9	>0.9	>0.9
aSPU	**>0.9**	**>0.9**	**>0.9**	**>0.9**

Experiment 4B: Healthy vs. Suspicious Dispersions

In the second part of the experiment, the dispersions of each population were compared using the hypothesis test described in [71]. The null hypothesis for this experiment is that the dispersions of each population are the same. The alternative hypothesis for this experiment is that the dispersions of each population are different.

Table 4.18: Results (p-values) for healthy scan versus suspicious scan dispersions in the 2014 phantom data set.

Test	STFT	DFT	EMD	TDF
Permuted	>0.9	0.8	>0.9	0.6
Observed	>0.9	0.8	>0.9	0.6

Discussion

None of the experiment results were significant. This is unexpected considering that the 2017 clinical data set results were more significant in several cases. These results could be due to the overall low signal amplitude of the data set. Consistently low amplitude signals could be an indication of a low SNR. Therefore it may be difficult to extract useful information from the signals in this data set. This may be an indicator for the next system prototype that, within safety margins, higher power may be used to increase the signal SNR and ultimately improve the system's anomaly detection capabilities.

4.2.5 Experiment 5: Effect of Incorrect Signal Windowing

During initial observations of the signals of the 2017 clinical data set, the conclusion was incorrectly drawn that the first signal pulse (see Figure 3.1) was the segment of the signal that contained the most useful information. However, before this error was corrected several tests similar to Experiments 1, 2 and 3 were conducted using a windowing strategy that isolated that pulse. This section presents a sample of these results for comparison to the results produced by selecting the correct signal portion. The reason this comparative data is presented is that, in general, groups that study microwave radar for breast cancer detection all appear to face the challenge of correctly windowing the region of interest for further processing.

In this experiment, the means of the healthy and suspicious populations are compared using the adaptive sum of powers tests described in [70]. The null hypothesis for this experiment is that the means of each population are the same. The alternative hypothesis for this experiment is that the means of each population are different.

Table 4.19: Results (p-values) for healthy versus suspicious scan means in the 2017 clinical data set when incorrect windowing is used.

Test	STFT	DFT	EMD
SPU-1	0.6	0.5	0.7
SPU-2	0.7	0.7	0.8
SPU-3	0.9	>0.9	>0.9
SPU-4	0.6	0.7	0.6
SPU-5	>0.9	>0.9	>0.9
SPU-6	0.6	0.6	0.6
SPU-∞	>0.9	>0.9	>0.9
aSPU	>**0.9**	>**0.9**	>**0.9**

Discussion

The results of Experiment 5 are all insignificant. They suggest a higher degree of similarity between populations than the previous experiments. This is because, in this case, the portion of the signal that was used to generate the features did not contain any information about the subject being scanned. The initial pulse of the scan signals was determined to be strictly due to cross talk between the transmitting and receiving sides of the switching circuit as the pulse was being transmitted. Consequently there was no useful information the signals to begin with.

4.3 Machine Learning Experiment Results

4.3.1 Problem Statement

Our goal is to detect the presence of tumorous tissue in a human breast by analyzing a radio-frequency radar scan of the breast recorded using a hardware system like the one described in [46].

Each scan S_i comprises signals s_{ij}, recorded by an array of N_a transceiving antennas where $j \in \{1, ...N_a(N_a - 1)\}$. Each antenna transmits a pulse and the other antennas receive the signal, transmitted through and scattered by the breast tissue. Thus for each scan S_i, there are $N_a(N_a - 1)$ signals. We can extract a matrix of features X_i from each scan S_i which comprises vectors $\vec{x}_{ij}$ corresponding to each signal s_{ij} in the scan. Given a set of M scans represented by $N_a(N_a - 1)$ by N_s matrices, $\{S_{train1}, ...S_{trainM}\}$, which yield a set of $N_a(N_a - 1)$ by N_f feature matrices $\{X_{train1}, ...X_{trainM}\}$ and a set of labels $\{Y_{train1}, ...Y_{trainM}\}$ which indicate whether there was a tumor present in the breast tissue that produced each of these M scans, we would like to train a model that can predict the label Y_{new} of a new scan, S_{new} based on the feature array X_{new} generated from this new scan. The dimensions N_s and N_f depend on the the signal pre-processing and feature extraction methods used, respectively.

4.3.2 Receiver Operating Characteristic Evaluation

The ensemble classifier is trained using the $\hat{e}_{NP}$ metric, described in Section 3.5.4, Equation 3.21, to guide the selection of base models. The false-positive target rate, α, controls the weight placed on false-positive errors made by the base models and consequently, influences the receiver operating characteristic (ROC) of the overall ensemble classifier. By varying the α parameter over a range of values and evaluating the ensemble trained at each $\hat{e}_{NP}$ metric corresponding to each value, we can generate an ROC curve for the classifier. This allows differently configured or modified classifiers to be compared using the ROC curve's area-under-curve (AUC) value. However, in practice the curves produced by these experiments are not smooth or monotonic. Therefore, to get an ROC that can be more easily interpreted, the false-positive and true-positive values produced are first sorted by false-positive value, averaged if there are duplicate false-positive values and then linearly interpolated between the monotonic true-positive values observed and the conventional ROC curve endpoints (0,0) and (1,1).

4.3.3 Feature Comparison Experiment

In this section, the results of an experiment comparing the features proposed in Section 3.3 to the previously proposed PCA feature extraction method are presented. The feature extraction methods were compared using data from the 2014 clinical data set and the 2017 phantom data set. In this experiment, the predictive performance of the ensemble classifier was evaluated in terms of the AUC of the classifier's ROC curve. Two classification trials were considered equivalent if their AUC scores were within ±0.05 of each other since in this case it was found to be very difficult to distinguish between most ROC curves. The objective of this experiment was to test the following hypothesis.

Hypothesis

Performing a classification task using each of the features proposed in Section 3.3 will yield equivalent or better performance than using principal component analysis to generate features.

Results

The true-positive and false-positive rates for the ROC curve were generated by training and evaluating ensembles over 12 train-test folds from the 2014 clinical data set, where each test set comprised the scans from a single volunteer as in [46]. The 101 false-positive target rates used in [46] were used. In addition, the median peak amplitude threshold for antenna inclusion was set to 20 mV also as in [46], consequently the classifier only used features from the 185 strongest antenna pairs. Figure 4.27 shows the ROC curves for classification trials using each feature extraction method. Table 4.20 shows the AUC values corresponding to the curves in Figure 4.27.

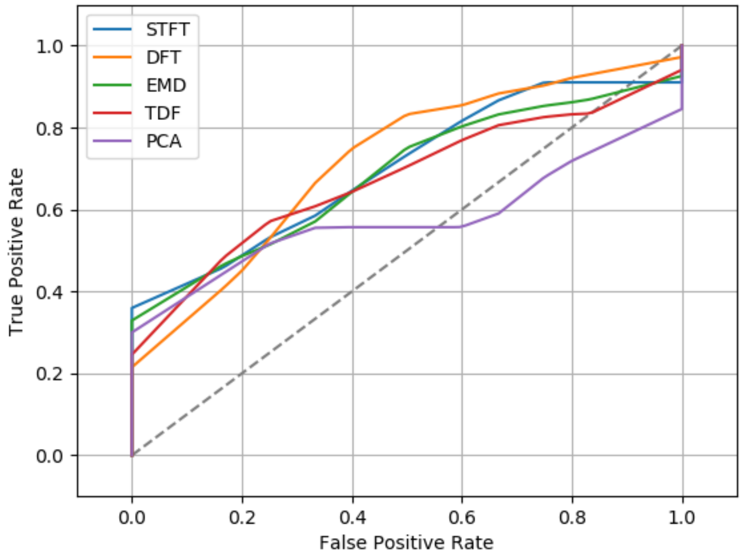


Figure 4.27: ROC curves produced by the ensemble classifier, when executed on 2014 clinical data set, using feature extraction methods described in Section 3.3.

Table 4.20: Area under curves for ROC curves shown in Figure 4.27. The actual AUC values are listed in the second column and the AUC values relative to the AUC yielded when PCA was used are listed in the third column.

Feature Extraction	Actual AUC	Relative AUC
PCA	0.58	-
STFT	0.70	+0.12
DFT	0.72	+0.14
EMD	0.69	+0.11
TDF	0.67	+0.09

Discussion

On the 2014 clinical data set, PCA performs notably worse than the other algorithms with only an AUC of 0.58. On the other hand, the performances of the other feature extraction methods were equivalent. Their AUC values were all within a 0.05 range. Therefore, the results of this experiment support the hypothesis.

These results suggest that PCA is not well suited for this data set. However, the results reported in [46] indicate significantly better performance using PCA features. This may be because the 2ν-SVM base models used in that implementation allow more fine-grained control over the false-positive and false-negative rates than the `scikit-learn SVC` [74] implementation used in this thesis. Since no other feature extraction methods are considered in [46], this neither supports or contradicts the hypothesis of this experiment.

The STFT features yielded the second best performance in this comparison experiment and represent an easily interpretable time-frequency decomposition of the radar signals. For these reasons, STFT features were used in the majority of the experiments in this section.

4.3.4 Hyperparameter Search Experiment

In this section, the hyperparameter search algorithms described in Section 3.5.2 are compared to the grid search (GS), used in the previous implementation, in the context of building model libraries. Specifically, the search algorithms are being evaluated as tools for rapidly finding good hyperparameter configurations to build the model library that the model ensemble is selected from. In this experiment, the predictive performance of the ensemble classifier was defined in terms of the average Neyman-Pearson score achieved by the classifier over all test sets. The runtime performance was defined in terms of the average time (in minutes) required to select and train an SVM ensemble. The objective of this experiment is to test the following hypotheses.

Hypothesis 1

Using each of the hyperparameter search methods proposed in Section 3.5.2 will yield equivalent or better predictive performance than using using an exhaustive grid search (over the hyperparameter grid in Table 3.4) to generate the model library used to build the ensemble.

Hypothesis 2

Using each of the hyperparameter search methods proposed in Section 3.5.2 will yield a faster average ensemble training time than using an exhaustive grid search to generate the model library used to build the ensemble.

Search Algorithm Configurations

In these experiments, the parameters of the search algorithms were tuned manually over several trial runs to prioritize exploration over exploitation since the goal in this experiment was to use the search algorithms to find a collection of good models rather than to simply find the best individual model. Ideally a more extensive investigation of parameter configurations would have been done, but due to time constraint, this was not possible. Each search algorithm in this experiment was given a budget of 1000 function evaluations. This number was selected to ensure the amount of computation preformed was reduced from the expected 4620 function evaluations performed by the grid search.

General For all search algorithms the maximum number of function evaluations was limited to 1000 evaluations. This was the only parameter that needed to be selected for the random search (RS) and random walk (RW) algorithms. The parameters selected for the other algorithms are described hereafter.

Simulated Annealing (SA) The initial temperature for this algorithm was set to 10,000 and the decay factor $\alpha = 0.98$. The temperature is initialized to this relatively high value to encourage exploration, especially in the early iterations of the algorithm. Similarly, the decay factor is set to a value very close to 1 so that the temperature decays gradually over the 1000 function evaluation budget.

Genetic Algorithm (GA) The mutation rate was set to 0.5. This means one half of the parameters in each chromosome were randomly modified. This is a relatively high mutation rate and was chosen to encourage exploration. The mutation factor b was set to 2.5. This value is quite a bit smaller than in [86], it was also chosen to encourage exploration since it essentially controls the decay of the chromosome mutations. The population size was 100. This population size was found to work well in practice.

Particle Swarm (PS) The individual coefficient c_1 was set to 3 and the social coefficient c_2 was set to 1, this was done to encourage particle to explore their local regions of the search space, rather than to rapidly converge to a single region. The inertia was set to a fixed value of $w = 0.5$. The particle population was set to 10. This population size was found to work well in practice.

Results

The Neyman-Pearson scores for ensembles generated over 12 train-tests folds from the 2014 clinical data set were computed. Each test set comprised the scans from a single volunteer as in [46]. A false-positive target rate of 0.05 was used in this experiment. Only a single value was investigated due to time constraint. In addition, since most of these search algorithms have some random component, the classification task was repeated 10 times for each search algorithm. The average $\hat{e}_{NP}$ values over the 12 test folds, from each of the 10 trials were used to generate box plots of the Neyman-Pearson scores. The median peak amplitude threshold for antenna inclusion was set to 20 mV also as in [46], consequently the classifier only used features from the 185 strongest antenna pairs. The STFT feature extraction method was used. Each trial was run on a computing cluster with the following resources allocated: 10 Intel Xeon CPUs @ 2.7 GHz and 50 GB of RAM. The hyperparameter searches for each antenna pair were performed asynchronously (in a non-serial order) to take advantage of the computing resources.

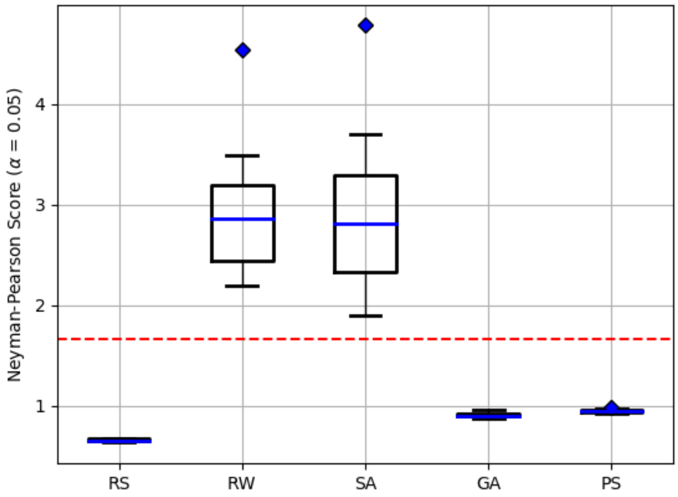


Figure 4.28: Boxplot of average Neyman-Pearson scores over 10 trials for each hyperparameter seach method. Outliers are shown as blue diamonds. The dashed red line represents the Neyman-Pearson score achieved by the classifier when an exhaustive grid search is used.

Table 4.21: Average ensemble train time achieved using each of the hyperparameter search methods described in Section 3.5.2 to build the model libraries. The actual average training time is shown in the second column. The third column shows the time relative to the avarage training time when GS is used.

Search Method	Actual Time (mins)	Relative Time (mins)	Improvement Factor
GS	29.46	-	-
RS	7.33	-29.46	4.02
RW	7.61	-21.85	3.87
SA	7.73	-21.73	3.81
GA	7.63	-21.83	3.86
PS	7.35	-21.11	4.00

Discussion

Figure 4.28 indicates that at a false-positive target rate of 0.05, Some of the search algorithms are consistently able to achieve even better predictive performance than the exhaustive grid search. Specifically, RS, GA and PS yielded consistently lower average $\hat{e}_{NP}$. Therefore, Hypothesis 1 is satisfied for these 3 search algorithms. On the other hand both RW and SA consistently produced larger $\hat{e}_{NP}$ scores than the exhaustive grid search.

Additionally, since the search algorithm parameters have not been explored more extensively, it is difficult to conclude whether certain algorithms are superior to others. These results should be considered an initial investigation.

Table 4.21 shows that, when a computational budget is enforced, Hypothesis 2 is satisfied for all search algorithms evaluated. All algorithms outperformed the GS by 21.73 minutes, (a factor of 3.81), or greater. This is because these algorithms only explore a fraction of the search space and therefore do not require as much computation as the GS. While each search algorithm only performs less than a quarter of the cross-validations that GS does, the additional overhead likely prevents the improvement factor from being larger.

When a classification task is performed over a larger range of false-positive target values, the advantage of using one of the hyperparameter search methods is lost. This is because each false-positive target rate requires a separate hyperparameter search to be executed. However, only a single exhaustive grid search needs to be performed for all false-positive target rates since its behavior is independent of the false-positive target rate.

4.3.5 Ensemble Model Library Experiment

In order to evaluate the performance of the ensemble selection method described in Section 3.5.3, only the best models from each antenna pair will be used to build the model library. This reduces the range of models available for ensemble selection significantly, but allows experiments over a range of false-positive target rates to be executed in a feasible amount of time. Because the number of models available for selection is reduced significantly, the change in performance relative to the previous selection method is evaluated in this section. In this experiment, the predictive performance of the ensemble classifier was defined in terms of the AUC of the classifier's ROC curve. Two classification trials were considered equivalent if their AUC scores were within ±0.05 of each other. The objective of this experiment is to test the following hypothesis.

Hypothesis

The predictive performance of the ensemble classifier when a model library comprising only the best models from each antenna pair is used is equivalent or better than the performance when a model library comprising all the models from each antenna pair is used.

Results

The true-positive and false-positive rates for the ROC curve were generated by training and evaluating ensembles over 12 train-tests folds from the 2014 clinical data set, where each test set comprised the scans from a single volunteer as in [46]. The 101 false-positive target rates used in [46] were used. In addition the median peak amplitude threshold for antenna inclusion was set to 20 mV also as in [46], consequently the classifier only used features from the 185 strongest antenna pairs. The STFT feature extraction method was used. Table 4.22 shows the AUC values for the curves in Figure 4.29.

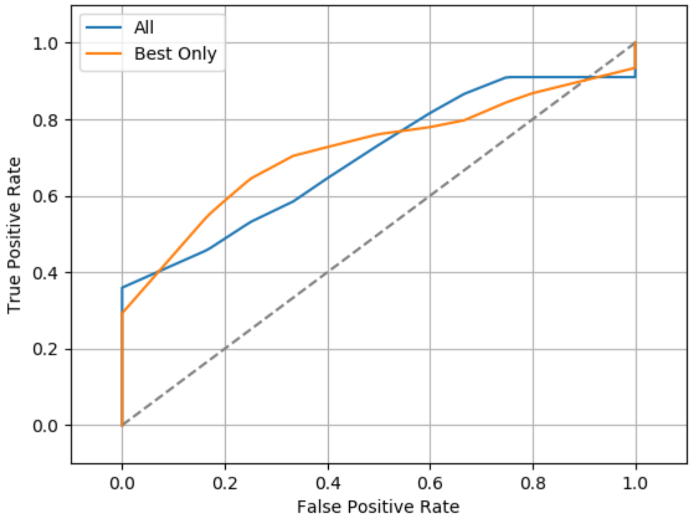


Figure 4.29: Receiver operating characteristic for 2014 clinical data using all models and only the best models from each antenna pair to build the model library.

Table 4.22: AUC values corresponding to ROC plots in Figure 4.29. The actual AUC values are listed in the second column and the AUC value relative to the AUC yielded when all models were used are listed in the third column.

Library Building Method	Actual AUC	Relative AUC
All Models	0.70	-
Best Models	0.72	+0.02

Discussion

The results shown in Figure 4.29 and Table 4.22 indicate that the predictive performance of the ensemble classifier when only the best models from each antenna-pair are used to build the model library is approximately equivalent to the performance when all models are used to build the model library.

4.3.6 Individual Hyperparameter Search Experiment

It is more effective to use an exhaustive grid search when performing a classification task over a range of false-positive target rates. If a classification task is performed at a fixed false-positive rate, then the results of Section 4.3.4 indicate that is much more efficient to use a hyperparameter search method instead. In addition, the results of Section 4.3.5 indicate that building model libraries using only the best models from each antenna pair is as effective as building the libraries using all hyperparameter configurations. The hyperparameter search methods can be used to find these models for each antenna pair. Therefore it is advantageous to determine which hyperparameter search method is able to find the best models. The hyperparameter search methods described in Section 3.5.2, with some of configurations described in Section 4.3.4, were compared to determine which method was best suited for this task.

Modified Configurations

The GA and PS algorithms were tuned to prioritize exploration over exploitation in Section 4.3.4. In this experiment we would like to simply determine which search methods is able to find the best individual hyperparameter configuration. Over several trial runs, this task was found to benefit from a more balanced configuration.

Genetic Algorithm The population size was 100 as before. The mutation rate was set to 0.25. This means one quarter of the parameters in each chromosome were randomly modified, as opposed to half in Section 4.3.4. This prevents more potentially good solutions from being lost due to random mutations. The mutation factor b was set to 5. This causes the aggressiveness to the mutations to decay faster also allowing good solutions found in the later stages of the search to be preserved more often.

Particle Swarm The particle population was set to 10, as before. The individual coefficient c_1 was set to 2 and the social coefficient c_2 was set to 2, the similarity between coefficient values provides a more balanced compromise between exploration and exploitation. The initial inertia was set to w = 1 and was linearly decayed at a rate of 0.01 per iteration. This results in a linear decay to 0 by the final function evaluation. This decay prevents good solutions found in the neighborhoods of particles in the later stages of the search to be ignored because particles have too much inertia.

Results

Each search method was limited to 1000 objective function evaluations. The objective function in this case was the average Neyman-Pearson score (Equation 3.21) yielded from cross validation of the model being evaluated over the available data. The data set used to evaluate these algorithms was the 2014 clinical data set. The STFT feature extraction method was used on the data set. The data set was partitioned into multiple train/test folds. The scans from each volunteer were grouped together as a test fold and the remaining scans corresponding to each set were used as training data. The best $\hat{e}_{NP}$ scores from each antenna pair were recorded then averaged to obtain a score for each false-positive target rate. The false-positive target rates used were 0.05, 0.1, 0.2 and 0.5. The median peak amplitude threshold for antenna inclusion was set to 20 mV also as in [46], consequently only features from the 185 strongest antenna pairs were considered. Figure 4.30 shows the results of this experiment. Only the genetic algorithm and the particle swarm algorithms preform better than a random search. The particle swarm algorithm is the best performing algorithm overall.

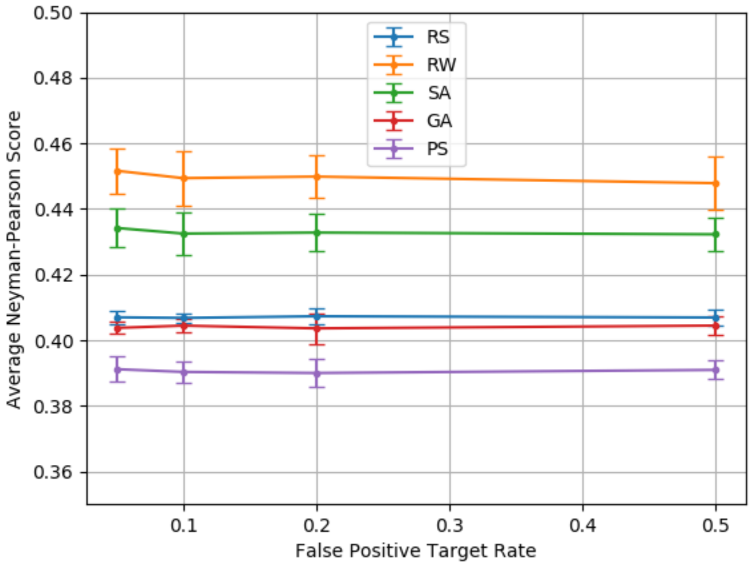


Figure 4.30: The best Neyman-Pearson ($\hat{e}_{NP}$) scores achieved by each hyperparameter search algorithm averaged over the 185 strongest antenna pairs of the 2014 clinical data set. The 95% confidence interval of each score is indicated by the vertical notched lines. Lower is better. The particle swarm algorithm (PS) performs the best. The other algorithms evaluated were random search (RS), random walk (RW), simulated annealing (SA) and genetic algorithm (GA).

Discussion

PS yielded the best performance followed by GA and RS. The weaker performance of SA and RW might be a consequence of the limited exploration allowed by these algorithms. Because these algorithms only allow limited exploration of the search space, they are very prone to being trapped in local optima. Since the PS algorithm was able to consistently produce the best models, this algorithm could be used to identify the models used to build the library when only the best model from each antenna pair is considered.

4.3.7 Ensemble Selection Experiment

In this experiment, the predictive performance of the forward stepwise selection (FSS) algorithm described in Section 3.5.3 is compared to ensemble selection using best individuals selection (BIS). In order for the experiment trials to be performed in a feasible amount of time, the following modifications were made to the original classification procedure described in Section 3.5.4. Firstly, the model library from which each ensemble was selected, comprised only the best models from each antenna pair. The impact of this change on the predictive performance is investigated in Section 4.3.5. Secondly, the ensemble size has been reduced to a total of 50 models (from 100, in Section 3.5.4), to further decrease the required computations. The predictive performance of the ensemble classifier was defined in terms of the AUC of the classifier's ROC curve. Two classification trials were considered equivalent if their AUC scores were within ±0.05 of each other. The objective of this experiment is to test the following hypothesis.

Hypothesis

Using each of the ensemble selection methods described in Section 3.5.3 will yield better predictive performance than selecting the best individual models to form the ensemble.

Results

The true-positive and false-positive rates for the ROC curve were generated by training and evaluating ensembles over 12 train-test folds from the 2014 clinical data set, where each test set comprised the scans from a single volunteer as in [46]. The false-positive target rates used in this experiment were {0.05, 0.1, 0.2, 0.5}. In addition the median peak amplitude threshold for antenna inclusion was set to 20 mV also as in [46], consequently the classifier only used features from the 185 strongest antenna pairs.

Four FSS configurations were evaluated. Firstly, the default FSS algorithm in which models are selected from the model library without replacement and the ensemble is initialized using the single best individual model (abbreviated, FSS-d). Secondly, FSS using selection with replacement and single model initialization was evaluated (FSS-r). Thirdly, FSS using selection without replacement and initialization using the 10 best individual models was evaluated (FSS-m). Finally, FSS using both selection with replacement and multiple model initialization was evaluated (FSS-rm).

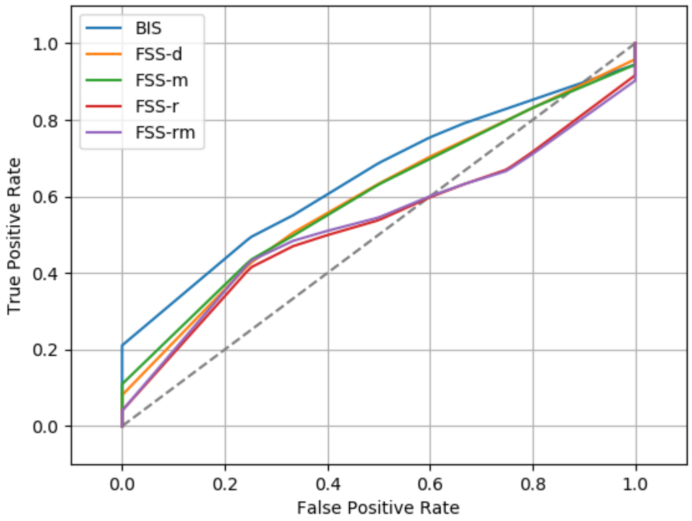


Figure 4.31: Receiver operating characteristic for 2014 clinical data using the BIS ensemble selection method and multiple configurations of the FSS method.

Table 4.23: Area under curve (AUC) values corresponding to ROC plots in Figure 4.31. The actual AUC values are listed in the second column and the AUC values relative to the AUC yielded the BM method was used are listed in the third column.

Ensemble Selection Method	Actual AUC	Relative AUC
BIS	0.65	-
FSS-d	0.60	-0.05
FSS-r	0.52	-0.13
FSS-m	0.60	-0.05
FSS-rm	0.53	-0.12

Discussion

In all cases, the FSS ensemble selection method yielded worse predictive performance than the BIS method. The FSS method performed better when replacement was not allowed. One possible explanation for poor performance of the FSS method is that the selection algorithm may have overfitted to the cross validation folds used to evaluate the model selections since the 2014 clinical data set is very limited (96 scans).

Chapter 5

Conclusions and Future Work

5.1 Conclusions

This thesis explored signal-processing approaches aimed at anomaly detection in data sets generated by a low-power microwave breast screening system. The algorithms developed were tested on limited data but are aimed to offer some preliminary directions for treatment of signals generated by the recently developed system.

5.1.1 Peak Absolute Voltage Analysis

The results in Section 4.1 demonstrate that, in general, the more recent data sets comprise much weaker signals than the previous data sets, this could contribute to the absence of statistically significant variation between features from tumor-bearing and tumor-free scans observed in Section 4.2, on the 2017 clinical and 2017 phantom data sets. The change in signal strength is due to significant changes in the hardware system since the 2014 clinical and 2014 phantom data sets were recorded.

5.1.2 Statistical Analysis Experiments

For the 2017 clinical data set, there is no statistically significant variation between the features of healthy and suspicious scans (with regard to distribution means and dispersion). Some significant results were obtained when the data set was partitioned by volunteer breast density. However, due to the small population size, especially when the data set was partitioned by volunteer breast density. This suggests that this radar screening system is better suited to monitoring breasts with higher density. It is not clear that any conclusions should be drawn based on these results. The statistically significant results observed in Section 4.2.1 may simply be due to random chance.

The results of Experiment 2 in Section 4.2.2 indicate that DFT features encapsulate differences in breast density well. As in Section 4.2.1, due to the small population sized used in this Experiment, there is a possibility that the results observed were influenced by random chance and do not represent a global trend.

In Experiment 3 (Section 4.2.3), features extracted from the 2014 phantom data set all demonstrate a statistically significant difference between tumor-bearing and tumor-free scans. However, in Experiment 4 (Section 4.2.4), such a difference is not observed in features extracted from the 2017 phantom data. Assuming the 2017 phantom data contains sufficient information for significant differences to be observed between populations, these results suggest that, different feature extraction methods are necessary to detect differences in the 2017 phantom data.

Experiment 5 in Section 4.2.5 demonstrated the importance of identifying the correct portion of the radar signals. Using the incorrect signal portion resulted in less statistically significant results than when the correct signal portion is used, (as in Experiment 1).

5.1.3 Machine Learning Experiments

The feature comparison experiment (Section 4.3.3) indicated that PCA does not work as well as the deterministic feature extraction methods (Section 3.3).

The results of the hyperparameter search experiment, presented in Section 4.3.4, show that using select hyperparameter searches to build a model library yields equivalent predictive performance to an exhaustive grid search, as well as faster training time per ensemble. Therefore, when performing a classification task at a fixed false-positive target rate, it is more effective to use a hyperparameter search method to build the model library. The random search algorithm in particular yielded superior predictive performance to the grid search.

The results of the model library experiment in Section 4.3.5 show that building model libraries using only the best models from each antenna-pair yields approximately equivalent predictive performance to building libraries from all hyperparameter configurations from all antenna pairs. Therefore, the classification task at a fixed false-positive target rate can be optimized by using only the best hyperparameter configurations from each antenna pair. Furthermore, these hyperparameter configurations can be identified more efficiently using a hyperparameter search method like particle swarm optimization.

The ensemble selection experiment in Section 4.3.7 demonstrated that using the FSS ensemble selection method yields worse predictive performance to ensemble selection using the BIS method. This was the case even when selection with replacement and multiple best model initialization was used.

5.2 Future Work

5.2.1 Feature Extraction

The DFT features in this thesis only include magnitude information from the scan signals. It is possible that the phase information also contains useful information that might further improve the classifier's predictive performance.

In addition, PCA is a powerful tool for reducing the dimensionality of data. It could be useful to use PCA in conjunction with the features proposed in 3.3 to yield low-dimensional feature vectors for each radar scan rather than just low-dimensional feature vectors for each scan signal. This could greatly simplify the classification algorithms used with out hurting the predictive performance.

5.2.2 Statistical Analysis

Repeating the experiments in Section 4.2 with a larger data set and consequently, higher experimental power is recommended to confirm the trends observed in the experiment results presented in this thesis.

5.2.3 Machine Learning

Many of the hyperparameter search algorithms used in this thesis have their own parameters that must be selected. Due to time constraint, most of the parameters were selected manually with relatively little investigation. More extensive tuning of these parameters could yield significantly different results.

Using an exhaustive grid search of hyperparameter is much more feasible when performing classification tasks over a range of false-positive target values because each model only needs to be evaluated once. On the other hand, in the hyperparameter-search-enabled classifier that was investigated in this thesis, a separate search is conducted for each false-positive target rate and the observed error rates of previously trained models are discarded. In the experiments presented in this thesis each search was limited to a maximum of 1000 hyperparameter configuration evaluations and the total size of the original hyperparameter grid was 4620. Consequently, the hyperparameter search classifier was likely to train and test many hyperparameter configurations redundantly. If the hyperparameter search classifier was updated to store the false-positive and false-negative rates yielded by each model evaluation and allowed these metrics to be re-used by future searches then the redundant training and testing would be eliminated. This, in turn, could make the hyperparameter search classifer more feasible for classification over a range of false-positive target rates.

The FSS ensemble selection method investigated in this thesis is only one ensemble selection that can be applied to this problem. Alternative selection methods may yield superior predictive performance. Some examples of algorithms that might be investigated are Pareto ensemble pruning proposed by Qian et al. [94], and GASEN proposed by Zhou et al. [95], boosting, bagging and stacking [72].

Finally, in this thesis, only ensemble support vector machines are used for generating models. Other machine learning algorithms such as neural networks and decision trees could be equally or more effective.

Bibliography

[1] Canadian Cancer Society. (2017). Breast cancer statistics, [Online]. Available: http://www.cancer.ca/en/cancer-information/cancer-type/breast/statistics/. Accessed on Sep. 17, 2018.

[2] ——, (2017). Prognosis and survival for breast cancer, [Online]. Available: http://www.cancer.ca/en/cancer-information/cancer-type/breast/prognosis-and-survival/. Accessed on Sep. 21, 2018.

[3] ——, (2017). What is breast cancer? [Online]. Available: http://www.cancer.ca/en/cancer-information/cancer-type/breast/breast-cancer/. Accessed on Sep. 17, 2018.

[4] Cancer Research UK. (2014). How can cancer kill you? [Online]. Available: https://www.cancerresearchuk.org/about-cancer/coping/physically/how-can-cancer-kill-you. Accessed on Sep. 21, 2018.

[5] H. Adami, D. Hunter, and D. Trichopoulos, *Textbook of Cancer Epidemiology*, ser. Monographs in epidemiology and biostatistics. Oxford University Press, 2008.

[6] Canadian Cancer Society. (2017). Mammography, [Online]. Available: http://www.cancer.ca/en/cancer-information/diagnosis-and-treatment/tests-and-procedures/mammography/. Accessed on Sep. 24, 2018.

[7] P. Hogg, J. Kelly, and C. Mercer, *Digital Mammography: A Holistic Approach*. Springer International Publishing, 2015.

[8] National Cancer Institute. (2015). Risk factors: Radiation, [Online]. Available: https://www.cancer.gov/about-cancer/causes-prevention/risk/radiation. Accessed on Sep. 24 2018.

[9] A. N. Sencha, E. V. Evseeva, M. S. Mogutov, and Y. N. Patrunov, *Breast Ultrasound*. Springer, Berlin, Heidelberg, 2013.

[10] Radiological Society of North America. (2018). Breast ultrasound, [Online]. Available: https://www.radiologyinfo.org/en/info.cfm?PG=breastus. Accessed on Sep. 24, 2018.

[11] A. Berger, "Magnetic resonance imaging," *British Medical Journal*, vol. 324, p. 35, Jan. 2002.

[12] M. J. Gabe, J. Boban, D. Djilas, V. Ivanovic, and H. Ojeda-Fournier, *Women's Imaging: MRI with Multimodality Correlation*. John Wiley & Sons, Mar. 2014, ch. BreastMRI: Introduction and Technique, pp. 239–264.

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

[14] M. Lazebnik, L. McCartney, D. Popovic, *et al.*, "A large-scale study of the ultrawideband microwave dielectric properties of normal breast tissue obtained from reduction surgeries," *Physics in Medicine & Biology*, vol. 52, no. 10, pp. 2637–2656, 2007.

[15] M. Lazebnik, L. McCartney, D. Popovic, *et al.*, "A large-scale study of the ultrawideband microwave dielectric properties of normal, benign and malignant breast tissues obtained from cancer surgeries," *Physics in Medicine & Biology*, vol. 52, no. 20, pp. 6093–6115, 2007.

[16] T. Sugitani, S. Kubota, S. Kuroki, *et al.*, "Complex permittivities of breast tumor tissues obtained from cancer surgeries," *Appl. Physics Lett.*, vol. 104, no. 25, pp. 3702–3707, 2014.

[17] R. C. Conceição, J. Jacob, and M. O'Halloran, *An Introduction to Microwave Imaging for Breast Cancer Detection*. Springer, Cham, 2016.

[18] E. C. Fear, "Microwave imaging of the breast," *Technology in Cancer Research & Treatment*, vol. 4, no. 1, pp. 69–82, 2005.

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

[20] A. W. Preece, I. Craddock, M. Shere, L. Jones, and H. L. Winton, “Maria m4: Clinical evaluation of a prototype ultrawideband radar scanner for breast cancer detection,” *Journal of Medical Imaging*, vol. 3, pp. 1–7, Jul. 2016.

[21] M. Klemm, I. J. Craddock, J. A. Leendertz, *et al.*, “Clinical trials of a uwb imaging radar for breast cancer,” in *Proceedings of the Fourth European Conference on Antennas and Propagation*, Apr. 2010, pp. 1–4.

[22] M. Donelli, I. J. Craddock, D. Gibbins, and M. Sarafianou, “A three-dimensional time domain microwave imaging method for breast cancer detection based on an evolutionary algorithm,” *Progress In Electromagnetics Research*, vol. 18, pp. 179–195, 2011.

[23] D. Byrne, M. Sarafianou, and I. J. Craddock, “Compound radar approach for breast imaging,” *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 1, pp. 40–51, Jan. 2017.

[24] E. C. Fear, X. Li, S. C. Hagness, and M. A. Stuchly, “Confocal microwave imaging for breast cancer detection: Localization of tumors in three dimensions,” *IEEE Transactions on Biomedical Engineering*, vol. 49, no. 8, pp. 812–822, Oct. 2002.

[25] B. R. Lavoie, M. Okoniewski, and E. C. Fear, “Estimating the effective permittivity for reconstructing accurate microwave-radar images,” *PLOS ONE*, vol. 11, no. 9, pp. 1–25, Sep. 2016.

[26] D. Kurrant, J. Bourqui, C. Curtis, and E. Fear, “Evaluation of 3-d acquisition surfaces for radar-based microwave breast imaging,” *IEEE Transactions on Antennas and Propagation*, vol. 63, no. 11, pp. 4910–4920, Nov. 2015.

[27] E. C. Fear, J. Bourqui, C. Curtis, D. Mew, B. Docktor, and C. Romano, “Microwave breast imaging with a monostatic radar-based system: A study of application to patients,” *IEEE Transactions on Microwave Theory and Techniques*, vol. 61, no. 5, pp. 2119–2128, May 2013.

[28] J. Bourqui, J. M. Sill, and E. C. Fear, “A prototype system for measuring microwave frequency reflections from the breast,” *Journal of Biomedical Imaging*, vol. 2012, pp. 1–12, 2012.

[29] M. O'Halloran, M. Glavin, and E. Jones, "Rotating antenna microwave imaging system for breast cancer detection," *Progress In Electromagnetics Research*, vol. 107, pp. 203–217, 2010.

[30] R. C. Conceição, D. Byrne, J. A. Noble, and I. Craddock, "Initial study for the investigation of breast tumour response with classification algorithms using a microwave radar prototype," in *2016 10th European Conference on Antennas and Propagation (EuCAP)*, Apr. 2016, pp. 1–2.

[31] R. C. Conceição, D. M. Godinho, D. Byrne, and I. Craddock, "Support vector machines to aid breast cancer diagnosis using a microwave radar prototype," in *2017 XXXIInd General Assembly and Scientific Symposium of the International Union of Radio Science (URSI GASS)*, Aug. 2017, pp. 1–3.

[32] A. H. Golnabi, P. M.Meaney, and K. D. Paulsen, "3d microwave tomography of the breast using prior anatomical information," *Medical Physics*, vol. 43, no. 4, pp. 1933–1944, 2016.

[33] N. Epstein, P.Meaney, and K. Paulsen, "3d parallel-detection microwave tomography for clinical breast imaging," *Review of Scientific Instruments*, vol. 85, no. 12, pp. 85–96, 2014.

[34] D. Tajik, F. Foroutan, D. S. Shumakov, A. D. Pitcher, E. A. Eveleigh, and N. K. Nikolova, "Real-time microwave imaging of breast phantoms with constrained deconvolution of planar data," in *2018 IEEE International Microwave Biomedical Conference (IMBioC)*, Jun. 2018, pp. 31–33.

[35] D. Tajik, D. S. Shumakov, A. S. Beaverstone, and N. K. Nikolova, "Quasi-real time reconstruction of the complex permittivity of tissue through microwave holography," in *2017 11th European Conference on Antennas and Propagation (EUCAP)*, Apr. 2017, pp. 3485–3488.

[36] D. Tajik, F. Foroutan, D. S. Shumakov, A. D. Pitcher, and N. K. Nikolova, "Real-time microwave imaging of a compressed breast phantom with planar scanning," *IEEE Journal of Electromagnetics, RF and Microwaves in Medicine and Biology*, vol. 2, no. 3, pp. 154–162, Sep. 2018.

[37] M. Asefi, M. OstadRahimi, A. Zakaria, and J. LoVetri, "A 3-d dual-polarized near-field microwave imaging system," *IEEE Transactions on Microwave Theory and Techniques*, vol. 62, no. 8, pp. 1790–1797, Aug. 2014.

[38] T. Reimer, M. S. Nepote, and S. Pistorius, "Initial results using an mlem-based reconstruction algorithm for breast microwave radar imaging," in *2018 18th International Symposium on Antenna Technology and Applied Electromagnetics (ANTEM)*, Aug. 2018, pp. 1–2.

[39] R. C. Conceição, M. O'Halloran, M. Glavin, and E. Jones, "Evaluation of features and classifiers for classification of early-stage breast cancer," *Journal of Electromagnetic Waves and Applications*, vol. 25, no. 1, pp. 1–14, 2011.

[40] M. O'Halloran, B. McGinley, R. C. Conceicao, F. Morgan, E. Jones, and M. Glavin, "Spiking neural networks for breast cancer classification in a dielectrically heterogenous breast," *Progress in Electromagnetics Research*, vol. 113, pp. 413–428, 2011.

[41] B. Gerazov and R. C. Conceicao, "Deep learning for tumour classification in homogeneous breast tissue in medical microwave imaging," in *IEEE EUROCON 2017 -17th International Conference on Smart Technologies*, Jul. 2017, pp. 564–569.

[42] R. C. Conceição, H. Medeiros, M. O'Halloran, D. Rodriguez-Herrera, D. Flores-Tapia, and S. Pistorius, "Initial classification of breast tumour phantoms using a uwb radar prototype," in *2013 International Conference on Electromagnetics in Advanced Applications (ICEAA)*, Sep. 2013, pp. 720–723.

[43] B. L. Oliveira, D. Godinho, M. O'Halloran, M. Glavin, E. Jones, and R. C. Conceição, "Diagnosing breast cancer with microwave technology: Remaining challenges and potential solutions with machine learning," *Diagnostics*, vol. 8, no. 2, May 2018.

[44] W. Sekkal, L. Merad, and S. M. Meriah, "A comparative study for breast cancer detection by neural approach for different configurations of the microwave imaging system," *Progress in Electromagnetics Research*, vol. 65, pp. 69–78, 2018.

[45] H. Song, Y. Li, and A. Men, "Microwave breast cancer detection using time-frequency representations," *Medical and Biological Engineering and Computing*, vol. 56, no. 4, pp. 571–582, Apr. 2018.

[46] Y. Li, E. Porter, A. Santorelli, M. Popović, and M. Coates, "Microwave breast cancer detection via cost-sensitive ensemble classifiers: Phantom and patient investigation," *Biomedical Signal Processing and Control*, vol. 31, no. 1, pp. 366–376, Jan. 2017.

[47] H. Kanj and M. Popovic, "A novel ultra-compact broadband antenna for microwave breast tumor detection," *Progress In Electromagnetics Research*, vol. 86, pp. 169–198, 2008.

[48] H. Bahramiabarghouei, E. Porter, A. Santorelli, B. Gosselin, M. Popović, and L. A. Rusch, "Flexible 16 antenna array for microwave breast cancer detection," *IEEE Trans. on Biomed. Eng.*, vol. 62, no. 10, pp. 2516–2525, Oct. 2015.

[49] L. Kranold, P. Hazarika, and M. Popović, "Investigation of antenna array configurations for dispersive breast models," in *Proc. 2017 11th Eur. Conf. on Antennas and Propagation (EUCAP)*, Mar. 2017, pp. 2737–2741.

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

[51] A. Santorelli, M. Chudzik, E. Kirshin, *et al.*, "Experimental demonstration of pulse shaping for time-domain microwave breast imaging," *Progress In Electromagnetics Research*, vol. 133, pp. 309–329, 2013.

[52] Y. Li, A. Santorelli, O. Laforest, and M. Coates, "Cost-sensitive ensemble classifiers for microwave breast cancer detection," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2015, pp. 952–956.

[53] Y. Li, A. Santorelli, and M. Coates, "Comparison of microwave breast cancer detection results with breast phantom data and clinical trial data: Varying the number of antennas," in *2016 10th European Conference on Antennas and Propagation (EuCAP)*, Apr. 2016, pp. 1–5.

[54] L. Kranold, M. Coates, and M. Popoviæ, “Variability in clinical data obtained with flexible time-domain radiofrequency breast monitor,” in *2018 18th International Symposium on Antenna Technology and Applied Electromagnetics (ANTEM)*, Aug. 2018, pp. 1–3.

[55] A. Santorelli, O. Laforest, E. Porter, and M. Popović, “Image classification for a time-domain microwave radar system: Experiments with stable modular breast phantoms,” in *2015 9th European Conference on Antennas and Propagation (EuCAP)*, Apr. 2015, pp. 1–5.

[56] B. T. Nicholson, A. P. LoRusso, M. Smolkin, V. E. Bovbjerg, G. R. Petroni, and J. A. Harvey, “Accuracy of assigned bi-rads breast density category definitions,” *Academic Radiology*, vol. 13, no. 9, pp. 1143–1149, 2006.

[57] E. Jones, T. Oliphant, P. Peterson, *et al.* (2001). SciPy: Open source scientific tools for Python, [Online]. Available: http://www.scipy.org/. Accessed on Jan. 26, 2019.

[58] I. Jolliffe, “Principal component analysis,” in *International Encyclopedia of Statistical Science*, M. Lovric, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 1094–1096.

[59] N. E. Huang, Z. Shen, S. R. Long, *et al.*, “The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis,” *Proc. of the Roy. Soc. of London A: Math., Physical and Eng. Sci.*, vol. 454, no. 1971, pp. 903–995, 1998.

[60] D. Laszuk. (2017). Python implementation of empirical mode decomposition algorithm, [Online]. Available: http://www.laszukdawid.com/codes. Accessed on Jan. 29, 2019.

[61] B. Hudgins, P. Parker, and R. N. Scott, “A new strategy for multifunction myoelectric control,” *IEEE Transactions on Biomedical Engineering*, vol. 40, no. 1, pp. 82–94, Jan. 1993.

[62] X. Chen and Z. J. Wang, “Pattern recognition of number gestures based on a wireless surface emg system,” *Biomedical Signal Processing and Control*, vol. 8, no. 2, pp. 184–192, 2013.

[63] L. Wasserman, *All of Statistics*. Springer-Verlag New York, 2004.

[64] S. E. Maxwell, H. D. Delaney, and K. Kelley, *Designing Experiments and Analyzing Data: A Model Comparison Perspective 3rd Ed.* Routledge, 2018.

[65] H. Hotelling, "The generalization of student's ratio," *The Ann. of Math. Statist.*, vol. 2, no. 3, pp. 360–378, Aug. 1931.

[66] Z. Bai and H. Saranadasa, "Effect of high dimension: By an example of a two sample problem," *Statistica Sinica*, vol. 6, no. 2, pp. 311–329, Apr. 1996.

[67] S. X. Chen and Y. L. Qin, "A two-sample test for high-dimensional data with applications to gene-set testing," *Ann. Statist.*, vol. 38, no. 2, pp. 808–835, Apr. 2010.

[68] M. S. Srivastava and M. Du, "A test for the mean vector with fewer observations than the dimension," *J. of Multivariate Analysis*, vol. 99, no. 3, pp. 386–402, Apr. 2008.

[69] T. Cai, W. Liu, and Y. Xia, "Two-sample test of high dimensional means under dependence," *J. of the Roy. Statist. Soc.: Series B (Statist. Methodology)*, vol. 76, no. 2, pp. 349–372, Aug. 2014.

[70] G. Xu, L. Lin, P. Wei, and W. Pan, "An adaptive two-sample test for high-dimensional means," *Biometrika*, vol. 103, no. 3, pp. 609–624, Sep. 2016.

[71] M. J. Anderson, "Distance-based tests for homogeneity of multivariate dispersions," *Biometrics*, vol. 62, no. 1, pp. 245–253, Mar. 2006.

[72] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning 2nd Ed.* Springer-Verlag New York, 2009.

[73] M. Davenport. (2005). The 2nu-svm: A cost-sensitive extension of the nu-svm, [Online]. Available: https://mdav.ece.gatech.edu/publications/d-tr-2005.pdf. Accessed on Aug. 3, 2019.

[74] F. Pedregosa, G. Varoquaux, A. Gramfort, *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[75] O. Sagi and L. Rokach, "Ensemble learning: A survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 4, pp. 1249–1267, 2018.

[76] L. M. Rios and N. V. Sahinidis, “Derivative-free optimization: A review of algorithms and comparison of software implementations,” *Journal of Global Optimization*, vol. 56, no. 3, pp. 1247–1293, Jul. 2013.

[77] P. Matuszyk, R. T. Castillo, D. Kottke, and M. Spiliopoulou, “A comparative study on hyperparameter optimization for recommender systems,” in *Workshop on Recommender Systems and Big Data Analytics (RS-BDA’16)*, 2016.

[78] J. Bergstra and Y. Bengio, “Random search for hyper-parameter optimization,” *J. Mach. Learn. Res.*, vol. 13, pp. 281–305, Feb. 2012.

[79] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, “Optimization by simulated annealing,” *Science*, vol. 220, no. 4598, pp. 671–680, 1983.

[80] M. Podolski and M. Rejment, “Scheduling the production of precast concrete elements using the simulated annealing metaheuristic algorithm,” *IOP Conference Series: Materials Science and Engineering*, vol. 471, p. 112 083, Feb. 2019.

[81] S. R. Kancharla and G. Ramadurai, “Simulated annealing algorithm for multi depot two-echelon capacitated vehicle routing problem,” in *Transportation Research Board 98th Annual Meeting*, Jan. 2019.

[82] J. H. Holland, *Adaptation in Natural and Artificial Systems.* The University of Michigan Press, 1975.

[83] J. McCall, “Genetic algorithms for modelling and optimisation,” *Journal of Computational and Applied Mathematics*, vol. 184, no. 1, pp. 205–222, 2005.

[84] Z. Zhou, F. Li, H. Zhu, H. Xie, J. H. Abawajy, and M. U. Chowdhury, “An improved genetic algorithm using greedy strategy toward task scheduling optimization in cloud environments,” *Neural Computing and Applications*, Mar. 2019.

[85] C. Yu, X. Yin, Z. Yang, and Z. Dang, “Sustainable water resource management of regulated rivers under uncertain inflow conditions using a noisy genetic algorithm,” *International Journal of Environmental Research and Public Health*, vol. 16, no. 5, 2019.

[86] Z. Michalewicz, *Genetic Algorithms + Data Structures = Evolution Programs.* Springer-Verlag, 1996.

[87] H. K. Lam, S. H. Ling, F. H. F. Leung, and P. K. S. Tam, "Tuning of the structure and parameters of neural network using an improved genetic algorithm," in *IECON'01. 27th Annual Conference of the IEEE Industrial Electronics Society*, vol. 1, Nov. 2001, pp. 25–30.

[88] R. Eberhart and J. Kennedy, "A new optimizer using particle swarm theory," in *MHS'95. Proceedings of the Sixth International Symposium on Micro Machine and Human Science*, Oct. 1995, pp. 39–43.

[89] Z. Zhang, J. Wang, H. Zhong, and H. Ma, "Research on residential load optimization model based on the adaptive harmony search-particle swarm optimization algorithm," *IOP Conference Series: Earth and Environmental Science*, vol. 238, pp. 12–20, Mar. 2019.

[90] Y. Xie, Y. Zhu, Y. Wang, *et al.*, "A novel directional and non-local-convergent particle swarm optimization based workflow scheduling in cloud–edge environment," *Future Generation Computer Systems*, vol. 97, pp. 361–378, 2019.

[91] R. Caruana, A. Niculescu-Mizil, G. Crew, and A. Ksikes, "Ensemble selection from libraries of models," in *Proc. 2004 21st Int. Conf. on Machine Learning*, Jul. 2004, pp. 18–26.

[92] R. Caruana, A. Munson, and A. Niculescu-Mizil, "Getting the most out of ensemble selection," in *Proc. 2006 6th Int. Conf. on Data Mining*, Dec. 2006, pp. 828–833.

[93] C. Scott, "Performance measures for neyman–pearson classification," *IEEE Transactions on Information Theory*, vol. 53, no. 8, pp. 2852–2863, Aug. 2007.

[94] C. Qian, Y. Yu, and Z. H. Zhou, "Pareto ensemble pruning," in *29th AAAI Conference on Artificial Intelligence*, 2015, pp. 2935–2941.

[95] Z.-H. Zhou, J.-X. Wu, Y. Jiang, and S.-F. Chen, "Genetic algorithm based selective neural network ensemble," in *Proceedings of the 17th International Joint Conference on Artificial Intelligence - Volume 2*, 2001, pp. 797–802.

Appendix A

2017 Clinical Data Set: Patient Information Recorded for Each Volunteer

- *Breast Size*: volunteer's bra size. One of: {A,B,C,D,DD}.
- *Age*: volunteer's age in years
- *Weight*: volunteer's weight in kilograms
- *Height*: volunteer's height in centimeters
- *Pre/Post-Menopausal*: whether volunteer is pre-menopausal or post-menopausal
- *Family History*: whether volunteer has a family history of breast cancer (yes or no). This suggests whether the volunteer might have a genetic disposition favorable to the growth of cancer.
- *Ultrasound Gel Used*: whether ultrasound gel was used as a matching medium (to fill the air gaps between the volunteer's breast and the antenna array embedded in the bra).
- *Breasts Scanned*: which of volunteer's breasts were scanned (left, right or both).
- *Breast Density*: tissue density level of volunteer's breasts using BI-RADS categorization [56]. One of: {1,2,3,4}.

- *Suspicion Present*: whether or not there is an anomaly present in the volunteer's breast tissue that warrants further medical investigation (yes or no). This particular datum represents the sort of departure from healthy baselines that the radio-frequency screening system is intended to detect.

- *Suspicion Location*: approximate location of anomaly present in volunteer's breasts.

- *Benign Tumor Present*: whether there is benign tumor tissue present in volunteer's breasts (yes or no).

- *Benign Tumor Location*: approximate location of benign tumor tissue.

- *Cyst Present*: whether there are one or more cysts present in the volunteer's breasts

- *Cyst Location*: approximate location of cyst(s).

- *Cancer Present*: whether there are cancerous tumors present in volunteer's breasts

- *Cancer Location*: approximate location of cancerous tumor(s).