J Digit Imaging (2013) 26:198–208
DOI 10.1007/s10278-012-9506-2

Feature Selection in Computer-Aided Breast Cancer Diagnosis via Dynamic Contrast-Enhanced Magnetic Resonance Images

Megan Rakoczy · Donald McGaughey · Michael J. Korenberg · Jacob Levman · Anne L. Martel

Published online: 25 July 2012

Abstract The accuracy of computer-aided diagnosis (CAD) for early detection and classification of breast cancer in dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI) is dependent upon the features used by the CAD classifier. Here, we show that fast orthogonal search (FOS), which provides a more efficient iterative manner of computing stepwise regression feature selection, can select features with predictive value from a set of kinetic and texture candidate features computed from dynamic contrast-enhanced magnetic resonance images. FOS can in minutes search candidate feature sets of millions of terms, which may include cross-products of features up to second-, third- or fourth-order. This method is tested on a set of 83 DCE-MRI images, of which 20 are for cancerous and 63 for benign cases, using a leave-one-out trial. The features selected by FOS were used in a FOS predictor and nearest-neighbour predictor and had an area under the receiver operating curve (AUC) of 0.889 and 0.791 respectively. The FOS predictor AUC is significantly improved over the signal enhancement ratio predictor with an AUC of 0.706 ($p=0.0035$ for the difference in the AUCs). Moreover, using FOS-selected features in a support vector machine increased the AUC over that resulting when the features were manually selected.

Keywords Breast cancer · Computer-aided diagnosis · Stepwise regression · Fast orthogonal search

M. Rakoczy
DLCSPM 4-5, National Defence,
101 Colonel By Dr.,
Ottawa, Canada K1A 0K2
e-mail: meganrakoczy@me.com

D. McGaughey (✉)
Department of Electrical and Computer Engineering,
Royal Military College of Canada,
Box 17000, Station Forces,
Kingston, ON, Canada K7K 7B4
e-mail: mcgaughey-d@rmc.ca

M. J. Korenberg
Department of Electrical and Computer Engineering,
Queen's University,
Kingston, ON, Canada K7L 3N6

J. Levman · A. L. Martel
Department of Medical Biophysics,
Sunnybrook Research Institute ,
2075 Bayview Avenue, Room S668,
Toronto, ON, Canada M4N 3M5

Introduction

Breast cancer is both the second leading cause of cancer deaths and the most common cancer diagnosed in women today. An improvement in the breast cancer survival rate has been attributed to earlier detection [1] and regular screening has been identified as essential for improving survival rates [2].

Mammography is the standard screening modality but recent prospective studies have shown that dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI) has a superior sensitivity when compared to mammography and ultrasound for the detection of breast cancer in women with a >25 % lifetime risk of breast cancer [3–5]. Furthermore, in women with recently diagnosed breast cancer, MRI can detect cancer in the contra-lateral breast that is missed by mammography [6, 7]. As a result of these studies, the American Cancer Society (ACS) has recently recommended that breast screening with MRI be used as an adjunct to mammography in all women with a 20 % or greater lifetime risk of breast cancer [8]. Sardanelli et al. [9] looked at data from five prospective screening trials and calculated a pooled sensitivity of 81 % and a positive predictive value of 53 % for DCE-MRI.

This indicates that although MRI is the most sensitive screening modality, there are still a significant number of women who undergo biopsy and are found to have a benign lesion.

Many groups have attempted to improve the accuracy of diagnosis by developing computer-aided diagnosis (CAD) systems for the classification of breast cancer from DCE-MRI. These use kinetic, spatial and texture features extracted from the images [10–14] as inputs into a classification routine. Hundreds of features can be calculated from DCE-MRI but the use of too many features with relatively small data sets is known to lead to overtraining of the classifier and poor performance with unseen data. Since these features can be highly correlated to each other, some combinations of features may add little predictive value to the classifier. The goal of feature selection is to choose features with the most predictive value from a large set of potential features. In CAD of breast cancer, several approaches have been used for feature selection including a stepwise method [15], automatic relevance detection [14] and genetic algorithms [16, 17]

One well-known feature selection technique is the stepwise regression algorithm [18] in which candidate features are used to create a functional expansion of a target vector. The feature vectors are fitted in order of significance in the reduction of the mean-squared error (mse) of the functional expansion. When a QR decomposition is used, the feature vectors are orthogonalized with respect to all the previously fitted vectors so that highly correlated features are not selected. Typically, implementation of this approach has not recognized that orthogonal functions computed for terms already in the model do not have to be recomputed after a new term is added to the model, nor does it use a quick method of measuring the benefit of adding a given candidate.

The fast orthogonal search (FOS) algorithm [19] can perform the same regression as the stepwise regression without explicitly computing the orthogonal feature set, yet still exploit advantages inherent in orthogonality. This results in the FOS algorithm being much more efficient in computation time and memory usage than the well-known stepwisefit algorithm as is shown in this paper.

FOS has been shown to be highly effective in selecting appropriate features to model biological applications [20, 21]. Shirdel et al. [21] used FOS to identify features that predict which patients are at high risk for neutropenia based on information collected in the first cycle of a six-cycle chemotherapy treatment. Minz and Korenberg [20] used FOS in the appropriate selection of motifs and interacting groups of motifs involved in gene regulation.

In this paper, we will describe how the FOS algorithm can be used in the automatic selection of features with predictive value from a large set of candidate features extracted from DCE-MRI images. The candidate feature set consists of 106 features extracted from a set of 83 DCE-MRI exams as well as the point-by-point cross-product of these features. The features selected by the FOS routine are used in a nearest neighbour (NN), support vector machine and FOS-based classifier to evaluate the predictive value of the features selected by FOS.

Materials and Methods

Feature Selection

CAD systems for the classification of breast cancer from DCE-MRI use a set of features computed from the series of images. The purpose of the FOS method is to search a large set of features (perhaps millions when cross-products are included) derived from DCE-MRI images and find a small set (two to six) of features that when used by a classifier have good classification performance. This is done without having a priori knowledge of which features would improve the accuracy and without bias as to which features have been used in previous studies.

The FOS algorithm is a modelling technique that determines the model terms of a functional expansion using an arbitrary set of non-orthogonal candidate functions [22, 23]. FOS creates the functional expansion in order of significance fitting the terms that reduce the mse the most first. Manually testing all the combinations of features of even a relatively small set of initial features may be computationally demanding due to the huge number of such combinations. FOS will automatically select the features or feature set that have predictive value.

The FOS algorithm [19, 22, 23] models a target signal $y(n)$ as a functional expansion given by

$$y(n) = \sum_{m=0}^{M} a_m p_m(n) + e(n) \quad (1)$$

where $p_m(n)$ are the selected model terms, a_m are their weights and $e(n)$ is the residual error. The terms $p_m(n)$ in the model are selected from a large set of candidate functions $p_c(n)$, $c=0, 1, \ldots, C-1$ in the order that reduces the mse of the functional expansion the most. An implicit orthogonalization carried out by the FOS algorithm ensures features selected have little common energy and thus additional terms added have additional energy (and predictive value) compared to the model terms already fit.

FOS implicitly uses Gram–Schmidt orthogonalization [19, 23] to determine orthogonal candidate functions which can also be used to obtain a functional expansion given by:

$$y(n) = \sum_{m=0}^{M} g_m w_m(n) + e(n) \quad (2)$$

where g_m are the weights of the orthogonal functional expansion and $w_m(n)$ are the corresponding orthogonal functions. The details of the FOS algorithm are given in Refs. [19, 22, 23].

One advantage of the FOS algorithm is that it does not need to explicitly compute and store the orthogonal functions $w_m(n)$ in Eq. 2. The well-known stepwise fit regression (SWF) algorithm [18] explicitly computes the orthogonal functions $w_m(n)$ in Eq. 2. Although the stepwise regression algorithm is typically computed using QR matrix decomposition, it can be iteratively computed in FOS. The MatLab *stepwisefit* function [24] always fits a constant (DC term) as its first term and normalizes the candidate features to be zero-mean unit energy. In this work, we verified that the FOS algorithm would fit the same features, in the same order, and with the same weights as the *stepwisefit* function in cases when FOS was forced to fit a DC term as the initial term in the functional expansion (in the results presented below, a DC initial term was not required in the FOS models identified). However, FOS is much more efficient computationally and in memory usage than the *stepwisefit* algorithm as the orthogonal functions $w_m(n)$ are not explicitly computed and stored. In addition, as the correlation values $D(m, r)$ are computed once and stored, FOS is not required to recompute point-by-point correlations involving the $w_m(n)$. Both the FOS and SWF algorithms are order MNC where M is the number of terms fitted, C is the number of candidate terms and N is the record length. Typically, the number of candidates C is very large and this term dominates the computation time. When both algorithms are compiled, the implicit orthogonalization in FOS makes it faster by avoiding the need to carry out a full linear regression C times whenever a new term is to be added to the model.

To use FOS for feature selection, the candidate functions $p_c(n)$ are the set of features from which a concise subset of features with predictive value are to be chosen. In some applications, FOS may force fit a constant term as the first term in the functional expansion but, instead, in the present work a constant term was included in the candidate feature set. The target function $y(n)$ has the value of A for cancerous lesions and B for benign exams, where the diagnosis has been clinically determined, and A and B can be arbitrarily set. The choice of A and B will affect the weighting of each class in the calculation of the mse. The candidate feature set $\{p_c(n)\}$ consists of kinetic features (based on the relative signal intensities) and texture features (based on the distribution of the intensities within the region of interest). These features are described in the "Features" section.

In addition to the original features derived from the DCE-MRI images, cross-products of the original features can be used as candidate features. Cross-products of features may have more predictive value than the original features and are thus worth testing in a feature selection algorithm [20, 25]. Employing the cross-product terms as predictors of cancer or benign status would probably not be obvious even to clinical diagnosticians. But this highlights the utility of FOS in that it is able to explore very large sets of candidate terms and within minutes find concise models with unobvious terms that serve as very good indicators of the status of a lesion.

The feature set described in the "Features" section is used as the original candidate set. Exhaustively creating point-by-point cross-product candidates from the initial candidate set results in a huge number of combinations of cross-product candidates. However, these cross-product candidate functions can be created as they are being tested as the next term in the FOS model and then discarded. There is no need to simultaneously store all the cross-product candidates in memory. This negates the need for a huge amount of memory to store the cross-product candidates. In addition, due to the computational efficiency of FOS, it can exhaustively search millions of candidate terms efficiently.

The cross-products are created using each possible combination, with repetition, of the initial feature set excluding constant (DC) terms. The Mth order cross-products are the product of M factors $p_{c_i}(n), i = 1, \ldots, M$ belonging to the initial feature set.

It has been noted that FOS can be biased when its features have dissimilar energy. Thus all the features to be described in the "Features" section will be normalized to have zero mean and unit energy before the FOS functional expansion is created. All initial features are normalized before the cross-product terms are created. In addition the features are normalized after removing the hold-out data in the leave-one-out trial, thereby not using information from the test data exam to normalize the rest of the training data set.

Assuming we have a feature $p'(n)$, the feature normalized to have zero-mean and unit energy is computed using

$$p(n) = (p'(n) - \mu)/\sigma \tag{3}$$

where μ is the sample mean and σ is the sample standard deviation of the training set data once the hold-out data is removed. Note that the normalized initial feature set computed using Eq. 3 is used to compute the cross-product features. These cross-product terms are not normalized again. The hold-out data in the leave-one-out trials is normalized to the same extent using the mean and standard deviation computed within the training set using Eq. 3.

The initial feature set $p_c(n)$ and the maximum order X_0 of cross-terms to include in the feature set are inputs to the FOS feature selection algorithm. The FOS algorithm will compute the cross-product features as required, negating the need to store all the cross-product features in memory at the same time. The FOS feature selection algorithm will return the selected features $p_m(n)$ and their weights a_m for the functional expansion in Eq. 1.

The features selected by FOS may be both initial features as well as cross-product features. The features are selected in order of significance, the first one fit, reduces the mse of the model the most. Since the FOS algorithm implicitly orthogonalizes the features with respect to the previously selected features, features that are highly correlated to previously fitted features will not be selected by FOS as they will not significantly decrease the mse of the functional expansion or add predictive value in the feature selection algorithm.

The features selected by FOS are then used by one of three different predictors to assess the predictive value of these features. The three predictors used are the NN predictor, a FOS-based predictor and a support vector machine classifier. Note that all the classifiers use the zero-mean unit energy initial candidate functions as calculated in Eq. 3 and also that these zero-mean unit energy initial candidate functions are used in computing cross-product terms.

MRI Protocol and Patient Cohort

The data set of 83 DCE-MRI breast exams used in this study was collected at the Sunnybrook Hospital as part of a screening trial conducted by Warner et al. [26]. The following research was approved by the Institutional Review Board of the Sunnybrook Hospital and the patients gave informed consent. Patients were between the ages of 25 and 65 and were at high risk of developing breast cancer (BRCA1 and BRCA2 mutation carriers). Imaging was carried out using a 1.5 T magnet (Signa, General Electric Medical Systems, Milwaukee, WI). Simultaneous sagittal imaging of both breasts using dual three-dimensional sagittal TR-interleaved SPGR sequences (TR/TE/flip angle, 18.4 ms/4.3 ms/40° from 28 partitions per breast) was performed both before and after a rapid intravenous injection of 0.1 mmol kg^{-1} of Gd-DTPA. Each volumetric bilateral acquisition was obtained in 2 min 49 s. Slice thickness was 2–3 mm, without a gap, using a matrix of 256×256 and a field of view of 18–20 cm. Each study comprised a baseline volume and four post-contrast volumes. Twenty cases with a malignant lesion (confirmed on histology) and 63 showing a focal benign lesion (either confirmed on histology or presumed benign after 2 years disease-free follow-up) were used in this paper.

The data was motion corrected using a 3-D non-rigid registration technique for breast images [27]. In addition to the five acquired volumes (referred to as the raw image $I_{raw}(x, y, n)$), enhanced images $I_{en}(x, y, n)$ and difference images $I_{dif}(x, y, n)$ were calculated. Equations for calculating the enhanced and difference images are in the Appendix. Note that enhanced image and difference images are constant for $n=0$ and thus these images were not used. Regions of interest (ROI) outlining each lesion were defined manually by thresholding the enhanced images. The exact edges of each lesion were not delineated. A hyper-intense subsection of the lesion was sampled to help ensure that the data we collect from malignant samples are not accidentally containing signals from healthy tissue immediately adjacent to the cancers we are studying.

Features

Features were computed from the raw DCE-MRI images, from enhanced images and from difference images. A set of 106 candidate features were computed from these three data sets. These features were normalized according to Eq. 3 and form the initial candidate set for the FOS feature selection algorithm.

It was noticed that some of the exams have a maximum signal pixel intensity as low as 40 and others as high as 1,000. Thus, for the raw and difference images, the variation in contrast between patients greatly affected the feature values derived from these images. Thus several features were divided by either the maximum pixel value in the ROI at all times or the maximum pixel value in the entire image at all times such that the features were invariant to the raw intensity level of the images [28]. This normalization is not required for the enhanced image set as it is already normalized relative to the pixel intensity at time $n=0$.

In Table 1, a summary of all features and which image set they belong to are shown. Notice, some of the features were computed for each time $n=0,\ldots,4$ or $n=1,\ldots,4$ depending if they are defined at $n=0$. Other features, such as the signal enhancement ratio (SER), have only one value for all times. The peak time of the time absorption curve (TAC) is another example of a feature in the original feature set. The features given in Table 1 were computed for a region of interest for each patient and a vector of that feature for a given time was used as a candidate function for FOS. A total of 106 features formed the initial feature set. Equations for the computation of the washout, derivative and SER are given in the Appendix.

Classifiers

The NN classifier [19] is a well-known technique to determine the class of a data sample given exemplar samples known to belong to the classes of interest. In breast cancer classification, the two classes are cancer and benign. For the NN classifier, the correlation coefficient (equation given in Appendix) between the test feature vector and the mean of the training feature vectors of each class is computed. The test data will be then classified as being in the same set as the training vector for which the correlation coefficient is the maximum. Using the mean of the training set vectors compensates for different numbers of training exemplars in the two classes; otherwise there is a tendency to favour the class having the most training samples. The zero-mean unit

Table 1 The 106 initial features

Features	Image sets		
	Raw	Enhanced	Difference
1. Constant (DC)	1		
2. Mean pixel value within ROI	5	4	4
3. Maximum pixel value over maximum pixel value in ROI at all times	4	4	4
4. Maximum Intensity Pixel at all times		1	1
5. Time of Maximum Intensity Pixel at all times		1	1
6. Maximum pixel value over maximum pixel value in entire image at all times	5	4	4
7. Maximum of the mean pixel intensity		1	1
8. Time of the Maximum of the mean pixel intensity		1	1
9. Maximum washout		1	1
10. Mean washout		1	1
11. Maximum derivative		3	3
12. Mean derivative		3	3
13. Total Pixels in ROI		1	
14. Percentage of Pixels that peak at n=4		1	
15. Percentage of Pixels that peak before n=4		1	
16. SER	1		
17. Standard Deviation	5	4	4
18. Skewness	5	4	4
19. Kurtosis	5	4	4
TOTAL FEATURES	106		

The values in the table indicate the number of features computed for each image set. Note the empty entries indicate that the feature was not computed for that image set

energy normalized candidate functions described in Eq. 3 were used by the nearest-neighbour classifier because if any feature is significantly larger than the others it will dominate the correlation coefficient calculation in the NN algorithm.

The average SER value was used in a threshold detector to classify the lesion. If SER≥1.1, the test data is classified as cancerous, otherwise it is classified as benign. The threshold of 1.1 was used by Hylton et al. [29] and Levman at al. [30].

In the feature selection process, cancerous lesions were given a value of A and benign lesions were given a value of B. Using the arrays of features extracted from the DCE-MRI images, a FOS model was created to model this sequence. The FOS algorithm returns the significant features $p_m(n)$ as well as the weights a_m for the functional expansion in Eq. 1. For the hold-out data, the significant features determined by FOS were normalized using the mean and standard deviation computed from the training set data and Eq. 3. Then, an output value $y(n)$ was computed using Eq. 1. This value $y(n)$ was used in a threshold detector, as shown in Fig. 1. If $y(n)$ was greater than or equal to zero then the test case was classified as cancerous, otherwise it was classified as benign. Note that a threshold other than zero may also be used in the detector.

The support vector machine (SVM) performs classification by determining a hyper-surface that separates the data into two categories [31]. This hyper-surface is determined by optimizing the distance between the two sets of data. A radial basis kernel function is added to assist in the separation of data.

Feature Validation

A leave-one-out trial was conducted to ensure the classifiers were not being trained and tested using the same data. In this study there were 83 sets of patient data of which 82 were used by the feature selection process and 1 was used as the hold-out test. The leave-one-out trial cycles through all 83 DCE-MRI exams, leaving out a different exam each time as the test exam. The features given in Table 1 were computed for each patient and a vector of that feature for the "training" patients was used as a candidate function for FOS. In addition to these initial candidates, FOS tested cross-product candidates up to the maximum order of cross-terms parameter as passed into FOS. The target output $y(n)$ for FOS is a vector consisting of values of A for cancerous exams and B for benign exams.

The FOS feature selection algorithm was run many times varying the maximum cross-term order X_0, the number of terms to be added in the FOS model M, and the target classifier vector (A, B) for cancerous and benign tumours. The maximum order of cross-terms X_0 was varied from 1 to 4. For 106 initial features, the number of features tested by FOS depending on X_0 is shown in Table 2. For X_0=4, over 5.5 million features are tested by the FOS algorithm. For the

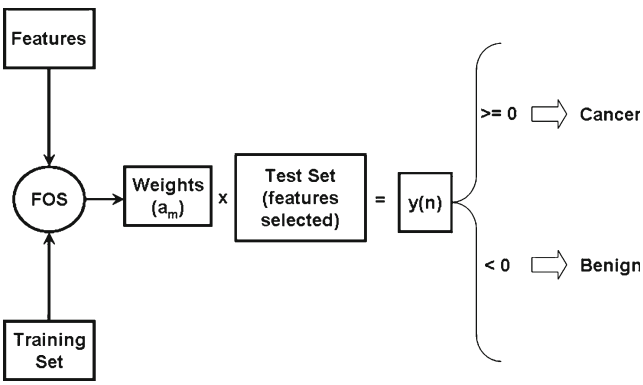

Fig. 1 When FOS is used as a classifier, the weights are multiplied with the test data exam of the features selected to provide a value for $y(n)$. If the value of $y(n)$ is greater than or equal to zero then the test case is classified as cancerous, otherwise it is classified as benign

Table 2 Memory requirements for the *stepwisefit* function and FOS algorithm

X_0	Number candidates	Stepwise fit memory	FOS memory (kB)
1	106	68.1 kB	68.1
2	5,671	3735 kB	68.73
3	204,156	129.28 MB	68.73
4	5,563,251	3,525 MB	68.73

X_0 order of cross-term candidates generated, *stepwisefit* Matlab function in MatLab Statistics Toolbox

MatLab *stepwisefit* algorithm, all the cross-term candidates are computed and passed into the *stepwisefit* function. However, for the FOS feature selection algorithm, the initial features are passed in and the cross-product features are computed as they are required, thus requiring only one more vector of 83 elements to store the cross-product features. The memory required to store the features is also shown in Table 2 for the *stepwisefit* MatLab© function and the FOS feature selector. Note, that for X_0=3, 4, the *stepwisefit* MatLab function was not able to execute as there was not enough memory (in a 2 GB system).

Our data set has 20 cancerous and 63 benign cases. Thus, using a target value of $(A, B) = (1, -1)$ results in the cancerous and benign cases respectively having average energies of 19/82 and 63/82 when a cancerous case is the hold-out data and 20/82 and 62/82 respectively when a benign case is the hold-out data As FOS minimizes the overall mse of the functional expansion, FOS will select features to model the benign cases over the cancerous cases since the benign cases have more energy. To give the benign and cancerous cases equal importance in minimizing the mse, (A, B) were chosen such that the cancerous and benign cases have the same average energy. This results in the target vector being set to $(A,B) = \left[\sqrt{63/19}, -1\right]$ when there are 63 benign and 19 cancerous lesions in the training data and $(A,B) = \left[\sqrt{62/20}, -1\right]$ when there are 62 benign and 20 cancerous lesions in the training data. Lastly, (A, B) were chosen such that the sum of target outputs for the cancerous and benign training cases were equal. This results in $(A,B) = [63/19, -1]$ or $(A,B) = [62/20, -1]$ for the above cases of benign and cancerous training data respectively.

In order to test the features selected by the FOS algorithm, the NN and FOS predictor were run on the hold-out data for all 83 iterations of the hold-one-out process. The area under the curve (AUC) for the receiver operating characteristic (ROC) curve was computed. For an ideal classifier, there is a threshold where the probability of true positives is 1 and the probability of false positives is zero. Thus, for an ideal classifier, the AUC is 1. For a random classifier the AUC is 0.5. However, a classifier could falsely predict all of a test set to be benign (and thus have a sensitivity of zero for predicting cancer) or all of the test set to be cancerous (and thus have a specificity of zero), yet still have an AUC=1 over the test set. In a clinical setting where novel cases are to be classified, a classifier that exhibits zero sensitivity or zero specificity over these cases has little value there, no matter how high its AUC. Thus, it is important to note the sensitivity, specificity and the significance level on Fisher's exact test, as well as reporting the AUC.

The AUC was also found for the SER predictor and a SVM predictor. The support vector machine employed a radial basis function kernel and utilizes a single parameter, γ, which controls the support vector radius. This parameter has the effect of controlling the tightness of the support vector machine generated classification function.

The SVM predictor was trained using the three features:

(a) The maximum signal intensity enhancement (as a percentage) from precontrast to any post-contrast image,
(b) Time of maximum enhancement in seconds and
(c) Maximum washout (as a percentage).

These are the features used in [30] and respectively correspond to feature 4 (enhanced image), 5 (enhanced image) and 9 (enhanced image) in Table 1. Each of these three variables were scaled from zero to one. For this SVM, γ=0.22.

A second SVM predictor was computed using features

1. The cross-product of the kurtosis in difference image at n=2 and the skew in the raw image at n=0
2. The cross-product of the time of the peak in the time absorption curves in the difference images and the mean pixel intensity in the enhanced images at n=1, and
3. The cross-product of the mean of the derivative of the enhanced image at n=4 and the maximum of the average intensity in ROI in enhanced images.

These features were selected using the FOS feature selector in a hold-one-out trial with only the initial feature set and cross-products up to second order (i.e. X_0=2), M=3 and (A, B)=(1, −1). The three features that were chosen the most often in the FOS model (70 out of 83 times) were each second-order cross-products, and were used to train this second SVM. In this trial, the SVM that used the cross-product features selected by FOS will be referred to as the FOS-SVM predictor. For this SVM, γ=0.136.

Results

Table 3 has the AUC for the NN and FOS predictor for the maximum cross-term order X_0=1, 2, 3, 4, terms fitted by FOS, the number of model terms M=3, 4, 5 and the target values (A, B)=(1, −1) and (A, B) chosen such that the benign and cancerous cases have equal energy and an equal sum in the training set. The two-tailed p value of the difference in the AUC

Table 3 The AUC for the NN and FOS predictors

X_0	M	Nearest neighbour			FOS predictor		
		(1, −1)	[1.82, −1] [1.76, −1]	[3.32, −1] [3.1, −1]	(1, −1)	[1.82, −1] [1.76, −1]	[3.32, −1] [3.1, −1]
1	3	0.699	0.699	0.773	0.694	0.702	0.752
	4	0.754	0.754	0.773	0.737	0.747	0.610
	5	0.757	0.757	0.754	0.591	0.598	0.640
2	3	0.616	0.423	0.739	0.602	0.660	0.889
	4	0.710	0.542	0.791	0.698	0.698	0.817
	5	0.694	0.612	0.783	0.615	0.699	0.772
3	3	0.562	0.638	0.710	0.438	0.653	0.646
	4	0.566	0.649	0.680	0.479	0.701	0.742
	5	0.608	0.690	0.75	0.514	0.706	0.660
4	3	0.319	0.493	0.533	0.272	0.398	0.621

X_0 order of cross-term candidates generated

M maximum number of features selected

[1.82, −1], [1.76, −1] column benign and cancerous cases in the training data have equal energy

[3.32, −1], [3.1, −1] column benign and cancerous cases in the training data have equal sums

Table entries are the AUC for that classifier

[32] was computed using ROC-KIT [33, 34]. For the AUC difference between the best FOS predictor (AUC=0.889) and the SER (AUC=0.706) the p value was 0.0035. The p value of the difference in the AUC for the best NN predictor (AUC=0.791) and the SER predictor (AUC=0.706) was 0.201.

The ROC for the best NN and FOS predictors in Table 3 is shown in Fig. 2 with the SER predictor ROC also plotted for comparison. Table 4 contains the sensitivity, specificity, positive predictive value and negative predictive value for the predictors in Fig. 2. For the FOS predictor a threshold of 0 was used and a threshold of 1.1 was used for the SER predictor.

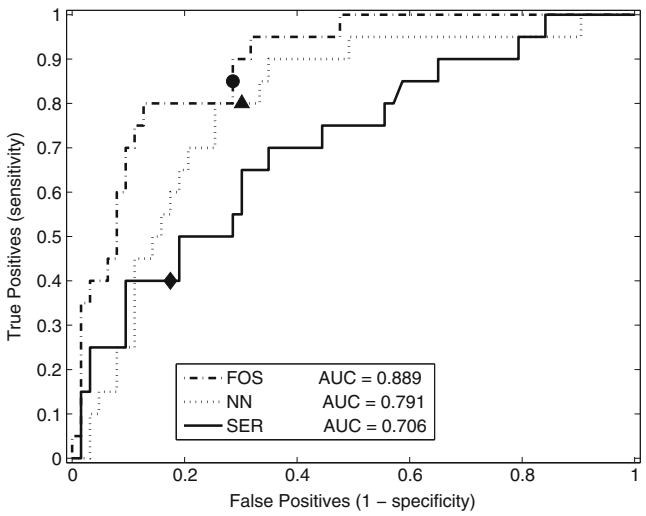

Fig. 2 The ROC curves with the highest area for the FOS and NN predictors from Table 3 are plotted with the ROC curve for the SER predictor for comparison. The AUC equals to 0.889, 0.791 and 0.706, respectively

The ROC for the SER, SVM and FOS-SVM predictors are shown in Fig. 3. The AUC is 0.706, 0.673 and 0.910, respectively. The p values of the differences in the AUC for the FOS-SVM predictor (AUC=0.910) and SVM predictor (AUC=0.673), and the FOS-SVM predictor (AUC=0.910) and the SER predictor (AUC=0.706), were 0.0060 and 0.0054 respectively. Table 5 contains the sensitivity, specificity, positive predictive value and negative predictive value for the predictors in Fig. 3. A threshold of 1.1 was used for the SER predictor. For the SVM, there is no typical threshold, so the sensitivity, specificity, positive predictive value and negative predictive value for the SVM and FOS-SVM predictors were computed at the same specificity (82.5 %) as the SER predictor.

Table 6 lists the features selected by the FOS feature selection algorithm and the number of times these features

Table 4 Hold-one-out Classifier Performance

Significance tests	FOS	NN	SER
Correct cancer diagnosis	17 of 20	16 of 20	8 of 20
Correct benign diagnosis	45 of 63	44 of 63	52 of 63
Sensitivity	85.0 %	80.0 %	40.0 %
Specificity	71.4 %	69.8 %	82.5 %
Fisher's exact two-tail test p	1.03×10^{-5}	1.1×10^{-4}	0.0634
Positive predictive value	48.6 %	45.7 %	42.1 %
Negative predictive value	93.8 %	91.7 %	81.3 %
Area under the ROC curve	0.889	0.791	0.706

FOS FOS classifier, *NN* nearest-neighbour classifier, *SER* signal enhancement ratio classifier

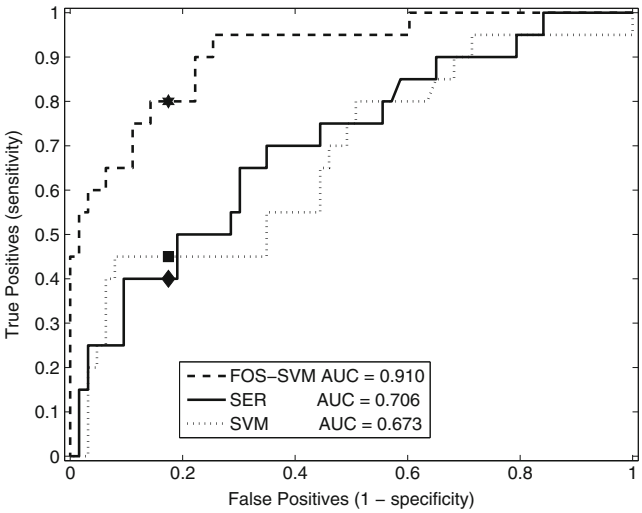

Fig. 3 The ROC curve is presented for the FOS-SVM, SER and SVM classifiers with the AUC equal to 0.910, 0.706 and 0.673, respectively

were selected in the hold-one-out trial for X_0=1 and M=2, 3, 4. The features that resulted in the FOS predictor with the highest AUC occur when X_0=2, M=3 and (A, B) is set so that the sum of the target outputs for benign and cancerous cases are equal. Table 7 lists the four most frequently selected features and describes the initial features used to create the cross-product features and the frequency these cross-product terms are selected. Table 8 lists the cross-product features that resulted in the NN predictor with the highest AUC.

Discussion

The individual features in Tables 7 and 8 that are selected most often to create the cross-products all seem to be reasonable choices. The moments of the image intensity values, i.e. standard deviation, skew and kurtosis, all provide information about the texture of the lesion, while the mean enhancement and time to peak intensity provide information about how much contrast arrives in the lesion and how

Table 5 SVM Classifier Performance

Significance tests	SVM	FOS-SVM	SER
Correct cancer diagnosis	9 of 20	16 of 20	8 of 20
Correct benign diagnosis	52 of 63	52 of 63	52 of 63
Sensitivity	45.0 %	80.0 %	40.0 %
Specificity	82.5 %	82.5 %	82.5 %
Fisher's exact two-tail test p	0.0177	6.15×10^{-7}	0.0634
Positive predictive value	45.0 %	59.2 %	42.1 %
Negative predictive value	82.5 %	92.9 %	81.3 %
Area under the ROC curve	0.673	0.910	0.706

SVM: SVM classifier using manually selected features, *FOS-SVM*: SVM classifier using features selected by FOS, *SER*: SER signal enhancement ratio classifier

Table 6 Frequency that features are selected by FOS in leave-one-out trial using 105 initial features only

M	Feature	(A, B)		
		(1, −1)	[1.82, −1] [1.76, −1]	[3.32, −1] [3.1, −1]
3	DC	83	83	0
	Mean pixel value at n=1 in enhanced image	80	80	80
	Maximum pixel at n=4 in enhanced image divided by maximum pixel in ROI	80	80	82
4	DC	83	83	0
	Number of Pixels in ROI	80	80	83
	Mean pixel value at n=1 in enhanced image	80	80	80
	Maximum pixel at n=4 in enhanced image divided by maximum pixel in ROI	82	82	82
	Maximum pixel at n=1 in difference image	0	0	62
	Kurtosis at n=2 in raw image	0	0	11
5	DC	83	83	0
	Number of pixels in ROI	83	83	83
	Maximum pixel at n=4 in enhanced image divided by maximum pixel in ROI	82	82	82
	Mean pixel value at n=1 in enhanced image	80	80	80
	Maximum pixel at n=1 in difference image	62	62	66
	Kurtosis at n=2 in raw image	11	11	16
	Kurtosis at n=4 in enhanced image	4	4	70

M maximum number of features selected, *[1.82, −1], [1.76, −1] column* benign and cancerous cases in the training data have equal energy, *[3.32, −1], [3.1, −1] column* benign and cancerous cases in the training data have equal sums DC is a constant candidate term with pm *(n)=1* for *n=0,..,N-1*

Table 7 The second-order cross-product features selected by the FOS in the hold-one-out trial for X_0=2, M=3 and the sum of (A, B) values set to be equal in the training set

	Feature 1	Feature 2	N
1	Kurtosis at n=2 in difference image	Skew in raw image at n=0	83
2	Time of peak in TAC in difference image	Mean pixel value at n=1 in enhanced image	71
3	Mean of the derivative of the enhanced image at n=4	Maximum of the average intensity in ROI in enhanced images	70
4	Standard deviation at n=4 in difference image	Maximum pixel in raw data at n=1 divided by the maximum pixel in the ROI at any time	10

N number of times selected in the 83 hold-one-out trials

Table 8 The second-order cross-product features selected by FOS in the hold-one-out trial for X_0=2, M=4 and the sum of (A, B) values set to be equal in the training set, which result in a NN classifier with a AUC of 0.791

	Feature 1	Feature 2	*N*
1	Kurtosis at n=2 in difference image	Skew in raw image at n=0	83
2	Time of peak in TAC in difference image	Mean pixel value at n=1 in enhanced image	71
3	Mean of the derivative of the enhanced image at n=4	Maximum of the average intensity in ROI in enhanced images	70
4	Skew at n=3 in difference images	Skew at n=4 in enhanced images	61
5	Standard deviation at n=2 in enhanced images	Skew at n=4 in enhanced images	15
6	Standard deviation at n=4 in difference image	Maximum pixel in raw data at n=1 divided by the maximum pixel in the ROI at any time	11
7	Mean of the derivative of the difference image at n=4	Mean pixel value at n=3 in enhanced image	8

Feature 1 first feature in the cross-product term, *Feature 2* second feature in the cross-product term, *N* number of times selected in the 83 hold-one-out trials

quickly it arrives. Many other studies have shown that these features can provide useful information in discriminating between malignant and benign lesions.

It can be seen in Table 3 that the highest AUC of 0.889 occurs for the FOS predictor for candidate sets involving up to second-order cross-products (X_0=2), M=3 and (A, B) set such that the sum of the desired outputs for cancerous and benign cases equals zero. The FOS predictor had a sensitivity of 85 % for predicting cancer and a specificity of 71.43 % (p= 0.0000103 on Fisher's exact two-tailed test). The highest AUC for the NN predictor is 0.791 for X_0=2, M=4 and (A, B) set such that the sum of the desired outputs for cancerous and benign cases equals zero. This NN predictor had a sensitivity of 80 % for predicting cancer and a specificity of 69.84 % (p=0.00011 on Fisher's exact two-tailed test). Note that as the number of terms M in the model increases, the AUC often decreases. This can be attributed to FOS over-fitting the training set so that the hold-out-set predictions are actually worse than before adding the additional feature(s).

From Fig. 3 and Table 5, it is clear that the FOS feature selection chose features (using the entire data set) with predictive value for the FOS-SVM predictor as the AUC increased from 0.673 for the original SVM to 0.910 for the FOS-SVM. The latter predictor had a sensitivity of 80 % for predicting cancer and a specificity of 82.5 % (p=6.15×10^{-7}) on Fisher's exact two-tailed test). Note that the FOS-SVM predictor was tested using a fixed set of three selected cross-product features in leave-one-out trials, instead of each time reselecting the features without reference to the held-out case. Thus direct comparison of the FOS-SVM predictor to the NN and FOS predictors in this paper is not possible.

Note that the FOS predictor has an AUC of 0.889, which is significantly improved over an AUC of 0.706 for the SER predictor where feature selection has not been used. The p value for the difference in the AUC for the FOS predictor and SER predictor is 0.0035. In fact all three predictors using features selected by the FOS feature selection algorithm (FOS, NN and FOS-SVM) have higher AUC than the SER predictor.

Conclusion

Using sets of DCE-MRI images from 83 patients, it was demonstrated that FOS was able to automatically select features that contain information to aid in the accurate classification of breast cancer. In addition to searching the 106 initial features derived from the DCE-MRI images, FOS was able to exhaustively search cross-products of these features up to the fourth order, while the stepwise fit algorithm could not run as there was not sufficient memory (2 GB). The FOS algorithm was used to select three features with predictive value which were then used to train the FOS-SVM predictor. The AUC for the FOS-SVM predictor was 0.910 and the difference in AUC between the FOS-SVM and SER predictors had a p value of 0.0054 which is highly significant.

In a second test, the FOS feature selection algorithm was used in a hold-one-out trial and the features were used in a FOS predictor and the NN predictor. The AUC for the FOS predictor was 0.889 and the difference in AUC between the FOS and SER predictors had a p value of 0.0035 which is highly significant.

The FOS feature selection algorithm was able to test over 5.5 million features (including cross-term features) in a leave-one-out test. FOS is able to explore very large sets of candidate terms and within minutes find concise models with unobvious terms that serve as very good indicators of the status of a lesion.

Acknowledgements The authors acknowledge Dr. Ellen Warner, Associate Scientist, Sunnybrook Health Sciences Centre, Toronto, ON, Canada, and the Canadian Breast Cancer Research Association for providing the DCE-MRI images used in this study. The authors thank the reviewers for their constructive comments on the manuscript.

Grants NSERC Discovery Grants, DRM and MJK

Appendix

The 5 acquired volumes are referred to as the raw image ($I_{raw}(x, y, n)$) and given by.

$$I_{raw}(x,y,n) = s(x,y,n),\ \mathrm{n} = 0,1,\ldots,4 \tag{4}$$

The enhanced images $I_{en}(x, y, n)$ and difference images $I_{dif}(x, y, n)$ were respectively generated by:

$$I_{en}(x,y,n) = \frac{s(x,y,n)}{s(x,y,0)},\ \mathrm{n} = 1,\ldots,4 \tag{5}$$

and

$$I_{dif}(x,y,n) = s(x,y,n) - s(x,y,0),\ \mathrm{n} = 1,\ldots,4 \tag{6}$$

where $s(x, y, n)$ represents the post-contrast image at time n, for $n=1,\ldots,4$ and $s(x, y, 0)$ is the precontrast image at coordinates (x, y).

The washout measures how quickly the contrast agent leaves the tissue and is computed on a pixel-by-pixel basis as

$$W.O. = \left(1 - \frac{I_{en}\ (x,y,4)}{max_{(x,y,n)}\ (I_{en}\ (x,y,n))}\right) \tag{7}$$

For the derivative, the sample difference is given by

$$\Delta(x,y,n) = \left(\frac{I(x,y,n+1) - I(x,y,n)}{t_{n+1} - t_n}\right),\ n = 1,2,3 \tag{8}$$

The maximum and mean W.O. and derivative in the ROI are used as features. The signal enhancement ratio (SER) was computed on a pixel-by-pixel basis given by the equation

$$SER = \frac{\max_{(x,y)}\ I_{raw}(x,y,1) - \max_{(x,y)}\ I_{raw}(x,y,0)}{\max_{(x,y)}\ I_{raw}(x,y,4) - \max_{(x,y)}\ I_{raw}(x,y,0)} \tag{9}$$

The correlation coefficient between vector x and y is given by

$$r_{\mathbf{xy}} = \frac{1}{P}\sum_{i=1}^{P}(\mathrm{x}_i - \overline{\mathrm{x}})(\mathrm{y} - \overline{\mathrm{y}})/\sigma_{\mathbf{x}}\sigma_{\mathbf{y}} \tag{10}$$

where x_i and y_i are the i^{th} elements, $\overline{\mathrm{x}}$ and $\overline{\mathrm{y}}$ are the sample means (of that vector's elements), and σ_{x} and σ_{y} are the sample standard deviations of x and y respectively.

References

1. Nass SJ, Henderson IC, Lashof JC, National Cancer Policy Board (U.S.). Committee on Technologies for the Early Detection of Breast Cancer., National Cancer Policy Board (U.S.). Committee on the Early Detection of Breast Cancer: Mammography and beyond: Developing technologies for the early detection of breast cancer. National Academy Press, Washington, DC, 2001
2. Curry SJ, Byers T, Hewitt M: Fulfilling the potential of cancer prevention and early detection: National Academies Press, 1 edition. (May 21, 2003), 2003
3. Warner E, Plewes DB, Hill KA, Causer PA, Zubovits JT, Jong RA, Cutrara MR, DeBoer G, Yaffe MJ, Messner SJ, Meschino WS, Piron CA, Narod SA: Surveillance of BRCA1 and BRCA2 mutation carriers with magnetic resonance imaging, ultrasound, mammography, and clinical breast examination. JAMA 292:1317–1325, 2004
4. Leach MO, Boggis CR, Dixon AK, Easton DF, Eeles RA, Evans DG, Gilbert FJ, Griebsch I, Hoff RJ, Kessar P, Lakhani SR, Moss SM, Nerurkar A, Padhani AR, Pointon LJ, Thompson D, Warren RM: Screening with magnetic resonance imaging and mammography of a UK population at high familial risk of breast cancer: a prospective multicentre cohort study (MARIBS). Lancet 365:1769–1778, 2005
5. Kriege M, Brekelmans CT, Boetes C, Besnard PE, Zonderland HM, Obdeijn IM, Manoliu RA, Kok T, Peterse H, Tilanus-Linthorst MM, Muller SH, Meijer S, Oosterwijk JC, Beex LV, Tollenaar RA, de Koning HJ, Rutgers EJ, Klijn JG: Efficacy of MRI and mammography for breast-cancer screening in women with a familial or genetic predisposition. N Engl J Med 351:427–437, 2004
6. Lehman CD, Gatsonis C, Kuhl CK, Hendrick RE, Pisano ED, Hanna L, Peacock S, Smazal SF, Maki DD, Julian TB, DePeri ER, Bluemke DA, Schnall MD: MRI evaluation of the contralateral breast in women with recently diagnosed breast cancer. N Engl J Med 356:1295–1303, 2007
7. Pediconi F, Catalano C, Roselli A, Padula S, Altomari F, Moriconi E, Pronio AM, Kirchin MA, Passariello R: Contrast-enhanced MR mammography for evaluation of the contralateral breast in patients with diagnosed unilateral breast cancer or high-risk lesions. Radiology 243:670–680, 2007
8. Saslow D, Boetes C, Burke W, Harms S, Leach MO, Lehman CD, Morris E, Pisano E, Schnall M, Sener S, Smith RA, Warner E, Yaffe M, Andrews KS, Russell CA: American Cancer Society guidelines for breast screening with MRI as an adjunct to mammography. Ca-a Cancer Journal for Clinicians 57:75–89, 2007
9. Sardanelli F, Podo F: Breast MR imaging in women at high-risk of breast cancer. Is something changing in early breast cancer detection? Eur Radiol 17:873–887, 2007
10. Lucht REA, Knopp MV, Brix G: Classification of signal-time curves from dynamic MR mammography by neural networks. Magn Reson Imaging 19:51–57, 2001
11. Gibbs P, Liney GP, Lowry M, Kneeshaw PJ, Turnbull LW: Differentiation of benign and malignant sub-1 cm breast lesions using dynamic contrast enhanced MRI. Breast 13:115–121, 2004
12. Schnall MD, Blume J, Bluemke DA, DeAngelis GA, DeBruhl N, Harms S, Heywang-Kobrunner SH, Hylton N, Kuhl CK, Pisano ED, Causer P, Schnitt SJ, Thickman D, Stelling CB, Weatherall PT, Lehman C, Gatsonis CA: Diagnostic architectural and dynamic features at breast MR imaging: multicenter study. Radiology 238:42–53, 2006
13. Vomweg TW, Buscema M, Kauczor HU, Teifke A, Intraligi M, Terzi S, Heussel CP, Achenbach T, Rieker O, Mayer D, Thelen M: Improved artificial neural networks in prediction of malignancy of lesions in contrast-enhanced MR-mammography. Med Phys 30:2350–2359, 2003
14. Szabo BK, Wiberg MK, Bone B, Aspelin P: Application of artificial neural networks to the analysis of dynamic MR imaging features of the breast. Eur Radiol 14:1217–1225, 2004
15. Chen WJ, Giger ML, Lan L, Bick U: Computerized interpretation of breast MRI: investigation of enhancement-variance dynamics. Med Phys 31:1076–1082, 2004

16. Mu T, Nandi AK, Rangayyan RM: Classification of breast masses using selected shape, edge-sharpness, and texture features with linear and kernel-based classifiers. J Digit Imaging 21:153–169, 2008
17. Sahiner B, Chan HP, Petrick N, Helvie MA, Goodsitt MM: Design of a high-sensitivity classifier based on a genetic algorithm: application to computer-aided diagnosis. Phys Med Biol 43:2853–2871, 1998
18. Draper NR, Smith H: Applied regression analysis. Wiley, New York, 1998
19. Korenberg MJ: Fast orthogonal algorithms for nonlinear system identification and time-series analysis. Advanced Methods of Physiological System Modeling 2:165–179, 1989
20. Minz I, Korenberg MJ: Modeling cooperative gene regulation using fast orthogonal search. The Open Bioinformatics Journal 2:80–89, 2008
21. Shirdel E, Korenberg M, Madarnas Y: Neutropenia prediction based on first-cycle blood counts using a FOS-3NN classifier. Advances in Bioinformatics 2011:1–8, 2011
22. Korenberg MJ: Fast Orthogonal Identification of Nonlinear Dif2ference Equation and Functional Expansion Models. Proceedings of Midwest Symposium on Circuits and Systems: Syracuse, NY 1987
23. Korenberg MJ: A robust orthogonal algorithm for system identification and time-series analysis. Biological Cybernetics 60:267–276, 1989
24. MatLab Statistics Toolbox 7.5 User's Guide, Nantick, MA: The MathWorks Inc., 2011
25. Shirdel E, Korenberg MJ, Madarnas Y: Use of fast orthogonal search to predict chemotherapy-induced leukopenia. Journal of Clinical Oncology, 2005 {ASCO} Annual Meeting Proceedings 23, 2005
26. Warner E, Plewes DB, Hill KA, Causer PA, Zubovits JT, Jong RA, Cutrara MR, DeBoer G, Yaffe MJ, Messner SJ, Meschino WS, Piron CA, Narod SA: Surveillance of BRCA1 and BRCA2 mutation carriers with magnetic resonance imaging, ultrasound, mammography, and clinical breast examination. JAMA: Journal of the American Medical Association 292:1317–1325, 2004
27. Martel AL, Froh MS, Brock KK, Plewes DB, Barber DC: Evaluating an optical-flow-based registration algorithm for contrast-enhanced magnetic resonance imaging of the breast. Physics in Medicine and Biology 52:3803–3816, 2007
28. Fienup JR: Invariant error metrics for image reconstruction. Applied Optics 36:8352–8352, 1997
29. Hylton N: Dynamic contrast-enhanced magnetic resonance imaging as an imaging biomarker. J Clin Oncol 24:3293–3298, 2006
30. Levman J, Leung T, Causer P, Plewes D, Martel AL: Classification of dynamic contrast-enhanced magnetic resonance breast lesions by support vector machines. IEEE Transactions on Medical Imaging 27:688–696, 2008
31. Vapnik V: The nature of statistical learning theory. Springer, New York, 2000
32. Hanley JA, McNeil BJ: A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 148:839–843, 1983
33. Metz CE, Herman BA, Roe CA: Statistical comparison of two ROC-curve estimates obtained from partially-paired datasets. Medical decision making: an international journal of the Society for Medical Decision Making 18:110–121, 1998
34. ROC-KIT Beta Version. Available at http://metz-roc.uchicago.edu/MetzROC. Accessed 30 April 2012