# Ensemble Learning for Breast Cancer Lesion Classification: A Pilot Validation using Correlated Spectroscopic Imaging and Diffusion-Weighted Imaging

Ajin Joy , Marlene Lin , Melissa Joines , Andres Saucedo , Stephanie Lee-Felker , Jennifer Baker , AiChi Chien , Uzay E Emir , Paul M. Macey , M Albert Thomas *

Posted Date: 8 May 2023

doi: 10.20944/preprints202305.0426.v1

Keywords: Correlated Spectroscopic Imaging; Diffusion weighted imaging; Machine Learning; Breast Cancer; Choline; Myo-inositol; Glycine; Water; Lipids

*Article*

# Ensemble Learning for Breast Cancer Lesion Classification: A Pilot Validation using Correlated Spectroscopic Imaging and Diffusion-Weighted Imaging

**Ajin Joy [1], Marlene Lin [1], Melissa Joines [1], Andres Saucedo [1,2], Stephanie Lee-Felker [1], Jennifer Baker [4], Aichi Chien [1], Uzay Emir [6], Paul M. Macey [5] and M. Albert Thomas [1,2,3,*]**

[1] Radiological Sciences,
[2] Physics and Biology in Medicine IDP,
[3] BioEngineering,
[4] Surgery of Nursing, University of California Los Angeles, Los Angeles, CA, United States
[5] School of Nursing, University of California Los Angeles, Los Angeles, CA, United States
[6] School of Health Sciences, College of Health and Human Sciences, Purdue University, West Lafayette, IN, United States
* Correspondence: Radiological Sciences, David Geffen School of Medicine at UCLA, 10945 Peter V Ueberroth Building, Suite#1417A, Los Angeles, CA 90095, Tel: (310) 206 4191, Fax: (310) 825 5837, Email: athomas@mednet.ucla.edu

**Abstract:** The main objective of this work was to evaluate the application of individual and ensemble machine learning models to classify malignant and benign breast masses using features from two-dimensional (2D) correlated spectroscopy spectra extracted from five-dimensional Echo Planar-Correlated Spectroscopic Imaging (5D EP-COSI) and Diffusion-weighted imaging (DWI). Twenty-four different metabolite and lipid ratios with respect to 2D diagonal peaks at 1.4ppm and 5.4ppm, and water from one-dimensional non-water-suppressed (NWS) spectra were used as the features. Additionally, water fraction, fat fraction and water-to-fat ratios from NWS spectra and apparent diffusion coefficients (ADC) from DWI were included. Nine most important features were identified using recursive feature elimination. XGBoost (AUC:93.0%, Accuracy:85.7%, F1-score:87.6%), GradientBoost (AUC:94.4%, Accuracy:87.0%, F1-score:89.4%), CatBoost (AUC:95.2%, Accuracy:86.9%, F1-score:88.4%) and RandomForest (AUC:92.2%, Accuracy:85.3%, F1-score:87.6%) were the best performing models. While the conventional biomarkers like choline, myo-Inositol, and glycine were statistically significant predictors, the key features contributing to the classification were ADC, 2D diagonal peaks at 0.9ppm, 2.1ppm and 2.3ppm, cross peaks between 1.4 and 0.9ppm, 4.3 and 4.1ppm, 2.1 and 1.4ppm and the triglyceryl-fat cross peak. The results highlight the contribution of the 2D spectral peaks to the model, and they demonstrate the potential of 5D EP-COSI for early breast cancer detection.

**Keywords:** Correlated Spectroscopic Imaging; Diffusion weighted imaging; Machine Learning; Breast Cancer; Choline; Myo-inositol; Glycine; Water; Lipids

## 1. Introduction

Breast cancer is one of the most prevalent cancers in females and one of the leading causes of cancer death worldwide (1, 2). Early detection and accurate characterization of breast malignancies are crucial factors in breast cancer management and positive treatment outcomes (3-12). Differentiation of benign from malignant breast lesions can aid clinicians in determining appropriate therapeutic plans. While histopathological examination of breast tissues extracted by biopsy is often required to confirm a suspicious lesion, a mammogram continues to be the gold standard for detection of breast cancer, but this approach has a high false positive rate (13). Multi-parametric MRI (mp-MRI), which includes dynamic contrast enhanced MRI (DCE-MRI), $T_2$-weighted MRI and diffusion-weighted imaging (DWI) may allow differentiating benign and malignant breast lesions

 doi:10.20944/preprints202305.0426.v1



that present highly overlapping enhancement patterns. However, despite the potential to eliminate unnecessary biopsies and follow-up examinations of benign tumors, mp-MRI-based breast tumor differentiation still has increased false positive findings.

Cell density, organization, membrane integrity and cellular metabolism of breast tissues undergo changes in the presence of cancer. Magnetic resonance spectroscopic imaging (MRSI) is capable of detecting the changes in concentrations of various metabolites and lipids in the tissue that are altered due to cancer related changes in cellular metabolism (14-23). High cell density and altered tissue structure due to cancer also lead to restricted motion of water molecules in the tissue, which can be measured by the apparent diffusion coefficient (ADC) on DWI (12, 24-31). DCE-MRI, one of the most sensitive diagnostic techniques, highlights the areas of increased blood flow and blood volume in the breast tissues due to cancer with the help of a contrast agent (5-7, 12, 32-35).

Even though the sensitivity of mp-MRI methods can be affected by various factors like tumor size and aggressiveness, these methods are often reported to have relatively high sensitivity (in the range of 88–100% for DCE-MRI, 85-95% for DWI and 80% for MRSI) (9, 12, 36-39). Reported specificity on the other hand, is relatively low (69-74% for DCE-MRI, 75–82% for DWI, and 74% for MRSI), restricting the capability in classification of benign and malignant lesions (37-40). While single-voxel spectroscopy has reported 64–82% sensitivity and 85–91% specificity (41), the multi-voxel technique of MRSI can cover a larger area of the breast with a relatively higher spatial resolution. Advanced MRSI techniques like five-dimensional (5D) echo-planar correlated spectroscopic imaging (EP-COSI) can record two-dimensional (2D) correlated spectroscopy (COSY) from multiple regions in three-dimensional (3D) space (42). Achieving high specificity is challenging due to overlapping patterns of diagnostic measures between benign and malignant lesions.

One option to potentially improve the specificity while retaining the benefit of the non-invasive nature of these imaging modalities is to use machine learning (ML) models to identify subtle or complex differences in the multi-model data that differentiate benign and malignant lesions (43, 44). Development and validation of machine learning models have seen an impressive growth in the last decade due to their high accuracy and flexibility in handling a wide range of datatypes and features (45). While individual machine learning models may perform well, a meta-approach that combines individual models named ensemble learning could generate even more generalizable models that can reduce individual base learner's variance or bias (46). In particular, advanced ensemble models like the gradient-boosted tree-based algorithm that combine multiple weak learners (decision trees) are shown to be capable of detecting key features of the multi-modal, multi-parametric imaging information for applications such as tissue/cancer grade classification (47-50).

Multiple studies have recently shown that the features extracted from DCE-MRI and DWI of breast tissues used in ML models are capable of predicting tumor grades, and classifying benign and malignant breast lesions (48-50). However, metabolite and lipid information from MRSI data has not been used in this context so far. Therefore, a major goal of this work was to evaluate the application of different machine learning models including ensemble learning techniques for the classification of benign and malignant breast lesions based on the 5D EP-COSI data with and without the corresponding ADC information from DWI data.

## 2. Materials and Methods

### *2.1. Subjects and Data Acquisition*

The dataset consisted of 5D EP-COSI and DWI data from twenty-three subjects with malignant breast masses (mean age 52 [range:33–71] years and seventeen benign breast masses (mean age 37 [range:19–60] years). All scans were acquired on a Siemens 3T Skyra scanner (Siemens Healthineer, Erlangen, Germany). The 5D EP-COSI data was acquired using FOV = 160×160×120 $mm^3$, matrix size = 16×16×8, TR/TE = 1500/35 ms, 64 $t_1$ points and 512 $t_2$ points with a spectral width of 1250 Hz and 1190 Hz along $F_1$ and $F_2$ respectively. A non-water-suppressed (NWS) 1D MRSI scan with one $t_1$ point was acquired for eddy current phase correction and phase correction for combining signals from multiple-receive coils (51). The data was non-uniformly undersampled (NUS) along two spatial $k_y$-

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 May 2023 doi:10.20944/preprints202305.0426.v1



$k_z$ and the spectral $t_1$ dimensions with a total acceleration factor of 8, and was reconstructed using Group Sparsity (GS) - based compressed sensing technique (52, 53).

The DWI acquisition protocol included the following: 2D spin-echo echo-planar imaging (EPI) sequence (TR/TE of 3800/93ms; data matrix, 192 × 192; signal average, 3; slice thickness, 3 mm; distance factor, 20%) in the axial plane. Diffusion sensitizing gradients (DSG) in three orthogonal directions with b values of 50 and 800 s/mm$^2$ were applied. The ADC maps were created automatically by the in-line scanner software using the trace-weighted images with b values of 50 and 800 s/mm$^2$.

### *2.2. Pre-Processing*

Tumor-containing slices in the DWI were selected and the boundaries of the lesion were marked by a radiologist. ADC values were then extracted from this delineated region of interest (ROI). The MRSI data were interpolated by a factor of 2 and the slices containing the tumor were identified similar to DWI. Spectroscopic voxels within the delineated region were extracted and the metabolite and lipid ratios were quantified in these voxels as described in (42). All variables were standardized with z-score normalization (zero mean and unit standard deviation) and voxels containing outlier measurements were removed. For the variables that followed a normal distribution, outliers were identified as three standard deviations away from the mean. For other variables, previously reported ranges of metabolite and lipid ratios were used as a guideline for outliers (42).

### *2.3. Feature Extraction*

Ratios of 24 metabolites and lipids from from water-suppressed 2D spectra and 1D NWS spectra including choline (Cho), myo-Inositol+glycine (mI+Gly), Unsaturated fatty acid and Triglyceryl fat cross-peaks with respect to methylene fat, olefinic fat and water were estimated from the MRSI data. The full list of metabolites and lipids identified in the 2D correlated spectroscopy (COSY) and 1D NWS spectra are shown in Table 1. The list of 99 quantitative features for ML analysis included metabolite and lipid ratios from 2D spectra, ADC values and water fraction, fat fraction and water-to-fat ratios from NWS spectra. A representative 2D COSY spectrum with labeled metabolite and lipid diagonal and cross peaks along with corresponding ADC map is shown in Figure 1.

**Table 1.** Metabolites and lipids identified in the 2D COSY and 1D NWS spectra of breast tissue.

| 2D COSY | | | | 1D NWS | |
|---|---|---|---|---|---|
| Diagonal Peaks | | Cross-peaks | | Peak label | Locations ($F_2$) ppm |
| Peak label | Locations ($F_2$,$F_1$) ppm | Peak label | Locations ($F_2$,$F_1$) ppm | | |
| Methyl Fat (FMETD) | (0.9, 0.9) | CP1 | (0.9, 1.4) | Methylene Fat (FAT14_1d) | 1.4 |
| Methylene Fat (FAT14) | (1.4, 1.4) | CP2 | (1.4, 0.9) | Water (WAT_1d) | 4.7 |
| Methylene Fat (FAT21) | (2.1, 2.1) | CP3 | (1.6, 2.3) | Olefinic Fat (UFD54_1d) | 5.4 |
| Methylene Fat (FAT23) | (2.3, 2.3) | CP4 | (2.3, 1.6) | | |
| Methylene Fat (FAT29) | (2.9, 2.9) | CP5 | (2.1, 1.4) | | |
| Choline (Cho) | (3.2, 3.2) | CP6 | (1.4, 2.1) | | |
| myo-Inositol+Glycine (mI+Gly) | (3.5, 3.5) | CP7 | (4.1, 4.3) | | |
| Methylene Glycerol Backbone (MGB41) | (4.1, 4.1) | CP8 | (4.3, 4.1) | | |
| Methylene Glycerol Backbone (MGB43) | (4.3, 4.3) | Unsaturated fatty acid cross peak, right lower (UFR_lower) | (2.1, 5.4) | | |
| Water (WAT) | (4.7, 4.7) | Unsaturated fatty acid cross peak, left lower (UFL_lower) | (2.9, 5.4) | | |
| Olefinic Fat (UFD54) | (5.4, 5.4) | Triglyceryl fat cross peak lower, (TGF_lower) | (4.2, 5.3) | | |
| | | Unsaturated fatty acid cross peak, right upper(UFR_upper) | (5.4, 2.1) | | |
| | | Unsaturated fatty acid cross peak, left upper (UFL_upper) | (5.4, 2.9) | | |
| | | Triglyceryl fat cross peak upper (TGF_upper) | (5.3, 4.2) | | |

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 May 2023 doi:10.20944/preprints202305.0426.v1



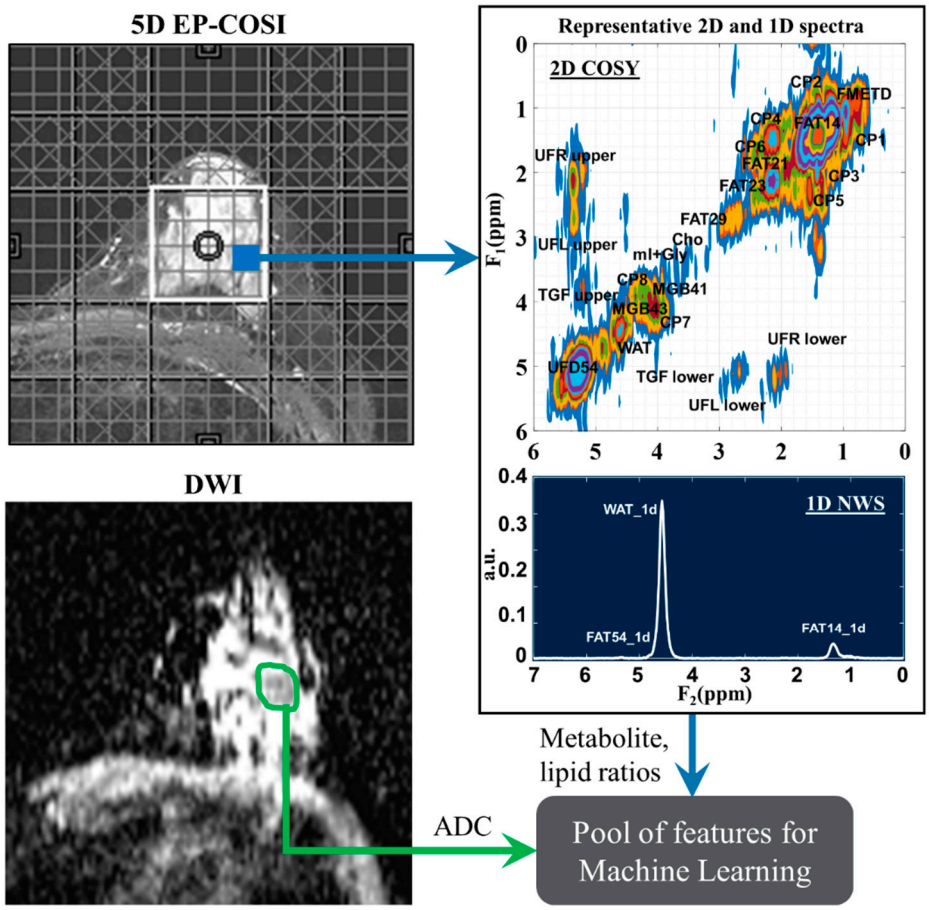



**Figure 1.** A representative 2D COSY spectrum and ADC map (age and diagnosis of this patient). Localizer image for 5D EP-COSI acquisition in shown on the top left panel. White box represents the placement of volume of interest. An extracted COSY spectrum and 1D NWS spectrum are shown to its right side. Bottom-left panel shows the corresponding ADC map for the same subject with region of lesion marked in green. These metabolite, lipid ratios and ADC values were inputted into the feature pool, which was then narrowed down using statistical tests and recursive feature elimination.

### *2.4. Feature Selection*

The entire dataset was split into approximately 80% for training and 20% for testing in such a way that the metabolite and lipid ratios from different voxels in the same lesion were not present in both training and testing sets. This ensured that the samples in the training and testing set were independent, which in turn avoided overestimation of model performance due to data leakage so that the model will be generalizable to new data.

Recursive feature elimination (RFE) was used to determine the best combination of variables to maximize the model's classification accuracy (54). A statistical significance test was used to narrow down the variable space before running the feature selection algorithms. Quantile-Quantile (Q-Q) plots of the data were used to check for normality and Levene's tests were used to check for the homogeneity of variances. Based on that, either a t-test or Mann-Whitney U (MWU) test for a statistical significance of p-value<0.05 was used. Correction for multiple comparisons was performed using Bonferroni correction methods. The base model for RFE was selected based on the model performance considering all the significant features identified in the statistical test.

### *2.5. Machine Learning Algorithms*

The open-source machine learning library for Python, 'scikit-learn' was used for implementing different supervised learning algorithms for classification (55), which included support vector

 doi:10.20944/preprints202305.0426.v1



machine (SVM), Naive Bayes, and K-nearest neighbors (KNN) as well as ensemble learning techniques including Adaptive Boosting (AdaBoost), GradientBoost, Extreme Gradient Boost (XGBoost), Light Gradient Boost, Categorical Boost (CatBoost) RandomForest and Decision Tree based bagging classifiers (56-58). In bagging, the training data was divided into different parts by random sampling with replacement and multiple models were trained on these different subsets. It then combined the prediction of each of the models by averaging. Boosting, on the other hand, used multiple base learners like decision trees in a sequential manner where the successive learner corrected for the error in prediction by previous one.

### *2.6. Cross-Validation and Parameter Tuning*

To ensure our classification model would generalize to unseen patient groups, the data was split into train and test sets in a way such that they were exclusive of each other - data from the same patient appears only in one set using the Shuffle Group Split. Likewise, the Grouped K-Folds cross validation method was used in both feature selection and hyperparameter tuning to return stratified folds with non-overlapping groups that are representative of the class distributions of the dataset. The train set was then standardized and the test set was standardized with the train set's statistics. The models were then optimized using the cross-validated Grid Search method, which exhaustively searched through the given hyperparameter space for the best combination of hyperparameters for the model and the training data.

### *2.7. Evaluation Metrics*

The classification performance of the different machine learning models in the testing stage was compared based on the scores of (a) accuracy (ratio of correct predictions to total number of predictions), (b) F1 score (2*((precision*recall)/(precision+recall))) and (c) area under the receiver operating characteristic (ROC) curve.

### *2.8. Statistical Analysis*

Statistical tests were performed to compare the performance of the machine learning models. One-way Analysis of Variance (ANOVA) test (in RStudio (version 4.1.1)) was used for this comparison based on the evaluation matrices for a statistical significance level of p-value<0.05. Tukey's HSD (honestly significant difference) post-hoc test was used for pair-wise analysis of these models with p-values adjusted for multiple comparisons using the Bonferroni method.

## **3. Results**

### *3.1. Feature Selection and Comparison*

Based on the results of the Mann-Whitney U test comparing the benign and malignant classes, the feature set was narrowed down to 86 that were statistically significant at $p \leq 0.05$. Nine out of these 86 features were identified as the most important by RFE, and included ADC, ratios of FAT21, FAT23, CP (1.4-0.9), CP (2.1-1.4), CP (4.3-4.1) and TGF upper with respect to FAT14, FMETD ratio with respect to WAT47 (1D) and the ratio of FAT21 with respect to FAT54 (1D). The boxplots of these most significant features are shown in Figure 2a for both malignant and benign classes. The values were z-score normalized. Figure 2b shows the correlation heatmap of these features.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 May 2023 doi:10.20944/preprints202305.0426.v1



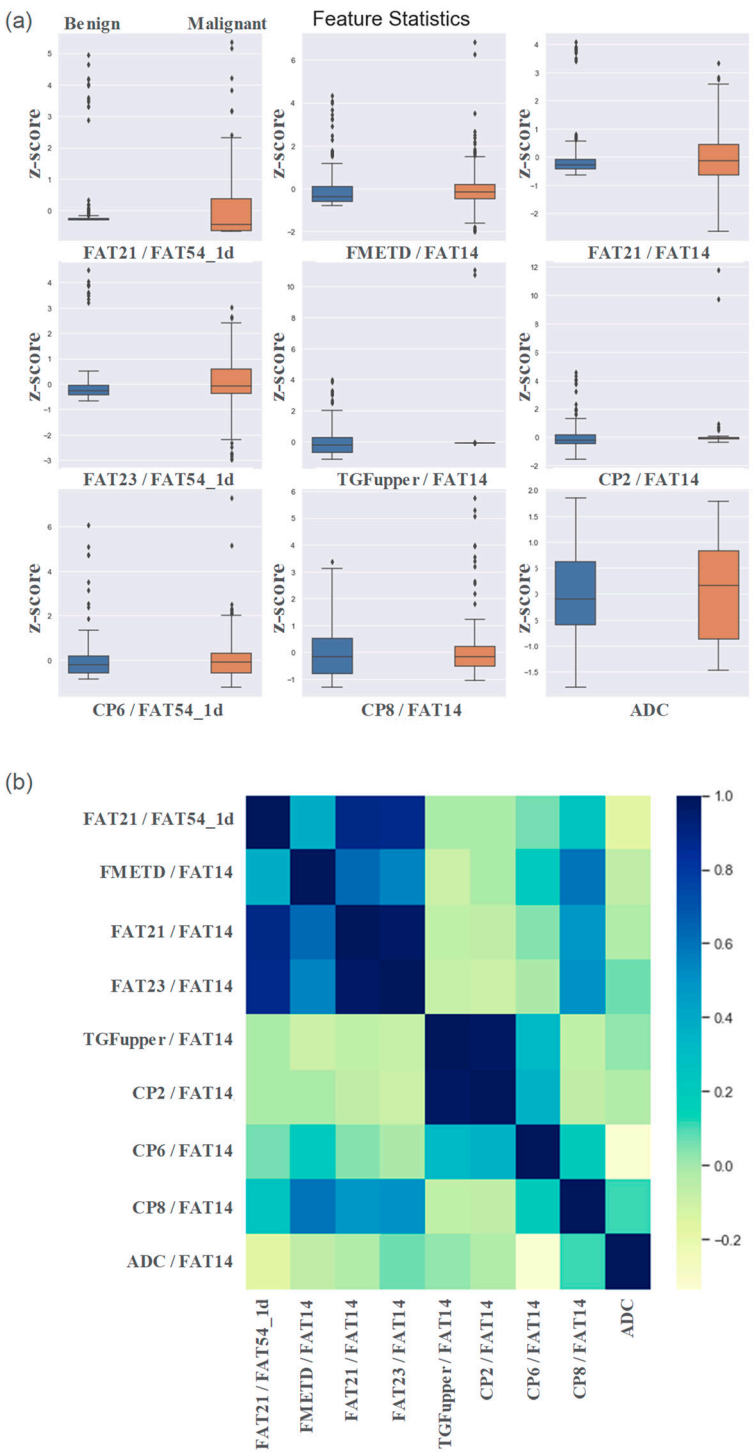



**Figure 2.** (a) Box plots of most important normalized features selected by recursive feature elimination process. (b) correlation heatmaps of the features.

### *3.2. Comparison of Models*

Comparative performance of linear SVM, DT-based bagging classifier, random forest, AdaBoost, GradientBoost, XGBoost and CatBoost are shown in Figures 3–5. These models were the best performing out of all the models considered in terms of performance metrics. Figure 3 shows the AUC, F1 score, and accuracy of these seven classifiers in the testing stage repeated 100 times with randomized dataset split and model initializations, and Figure 4 shows these scores in the cross-validation stage repeated 50 times. The respective box plots show the median and interquartile range

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 May 2023 doi:10.20944/preprints202305.0426.v1



(IQR) of these matrices, along with outliers. Their corresponding mean and standard deviation are listed in Table 2. ROC curves of these different models are shown in Figure 5.

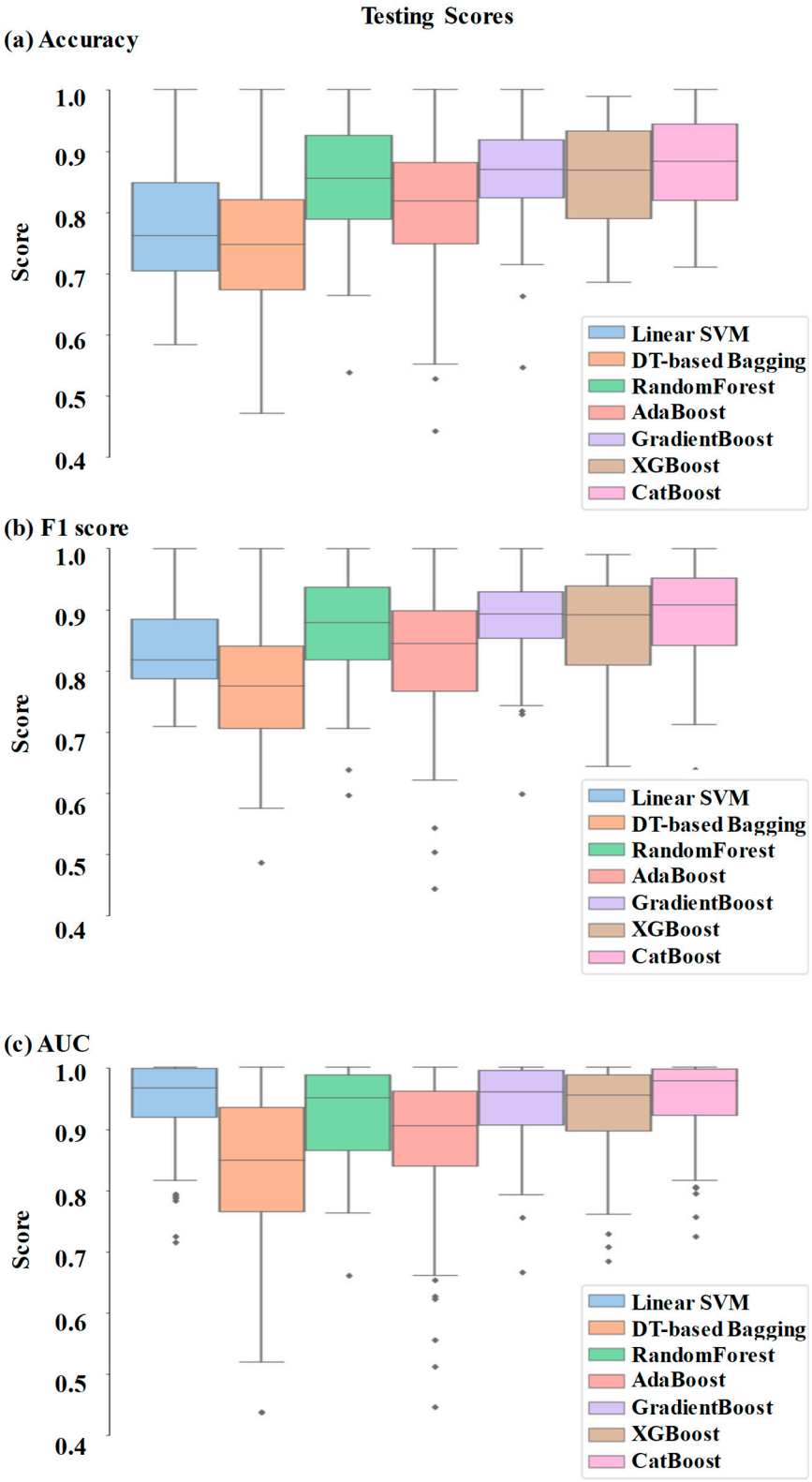



**Figure 3.** The box plots of (a) accuracy, (b) F1 Score and (c) AUC metrics of the ensemble model during the testing stage.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 May 2023 doi:10.20944/preprints202305.0426.v1



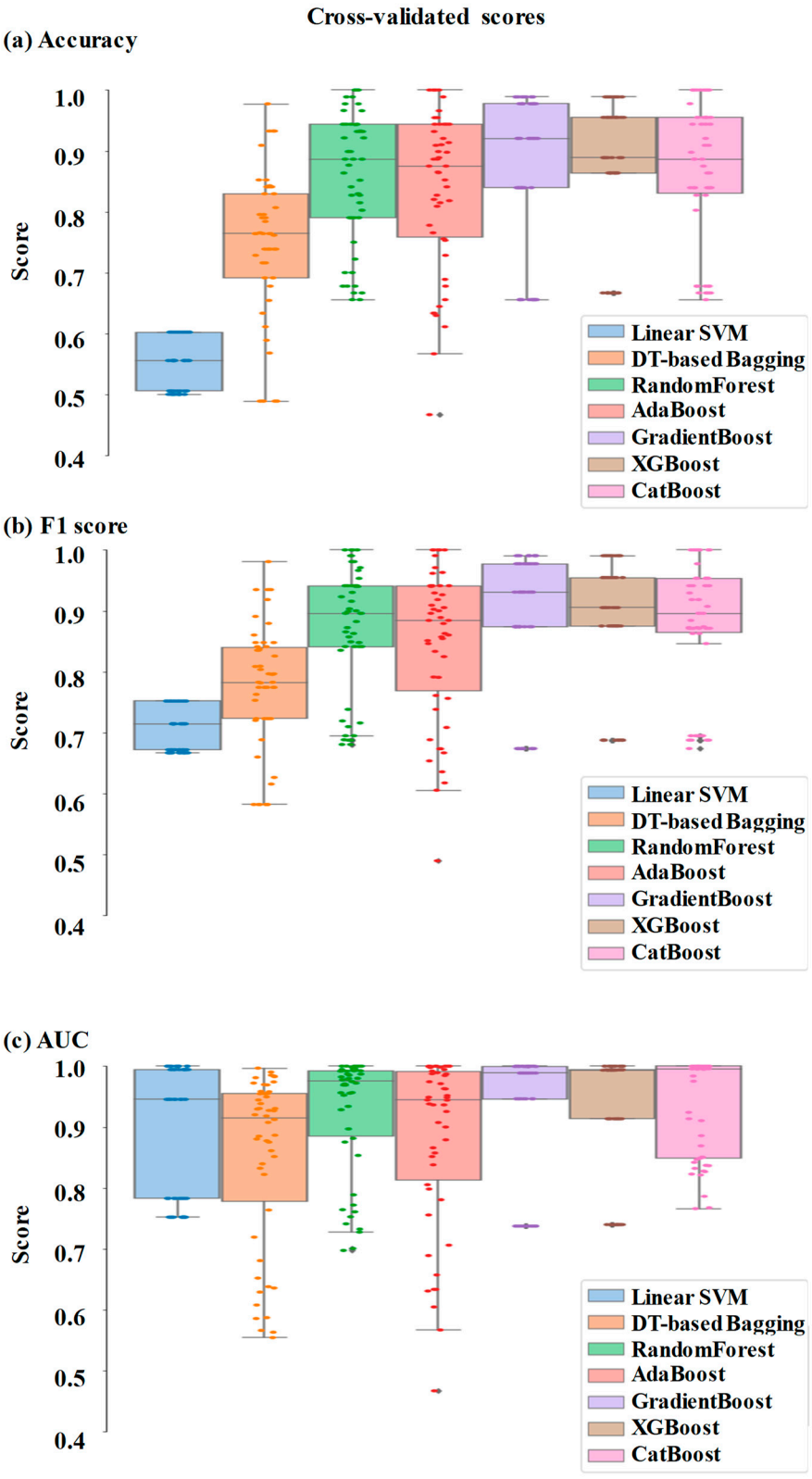



**Figure 4.** The swarm plots of (a) accuracy, (b) F1 Score and (c) AUC metrics of the ensemble model during the cross-validation stage.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 May 2023 doi:10.20944/preprints202305.0426.v1



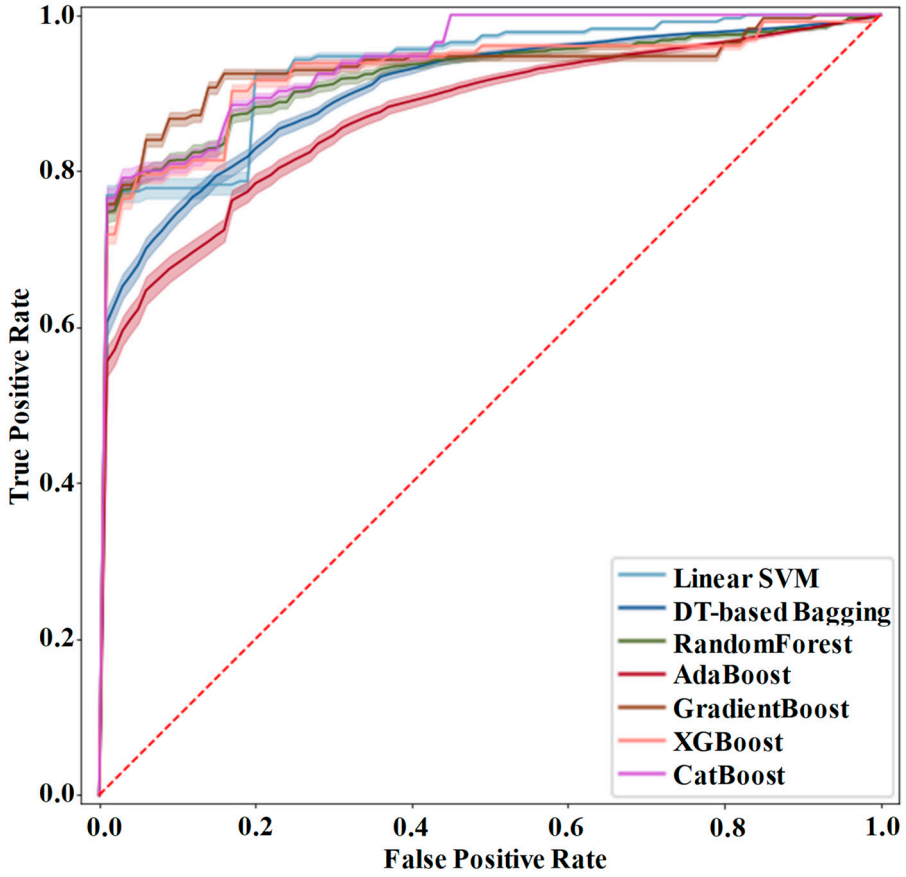



**Figure 5.** ROC curve of the ensemble models for differentiating malignant from benign breast tissues. The colored region surrounding the sloid line represents the standard error.

**Table 2.** AUC, Accuracy and F1 scores of ensemble models.

| Model | AUC (%) | Accuracy (%) | F1 score (%) |
|---|---|---|---|
| AdaBoost | 88.54 ± 10.79 | 81.08 ± 10.81 | 83.46 ± 10.39 |
| CatBoost | **95.17 ± 05.94** | 86.93 ± 09.58 | 88.44 ± 09.87 |
| DT-based Bagging | 84.08 ± 11.88 | 74.61 ± 11.09 | 78.62 ± 09.71 |
| GradientBoost | 94.37 ± 06.35 | **86.97 ± 07.90** | **89.39 ± 06.73** |
| Linear SVM | 94.77 ± 06.44 | 77.43 ± 09.86 | 83.94 ± 06.37 |
| RandomForest | 92.24 ± 07.30 | 85.31 ± 09.15 | 87.59 ± 08.06 |
| XGBoost | 92.98 ± 08.18 | 85.65 ± 09.61 | 87.64 ± 08.58 |

The results of Tukey's HSD post-hoc test following the ANOVA with p-values adjusted for multiple comparisons are shown in Table 3. It shows that the difference between the ensemble models XGboost, GradientBoost, CatBoost and RandonForest were not statistically significant in terms of Accuracy, AUC and F1 scores for the significance at $p \leq 0.05$. The feature importance and confusion matrix identified for these four classifiers are shown in Figures 6 and 7, respectively.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 May 2023 doi:10.20944/preprints202305.0426.v1



**Table 3.** Pairwise differences in AUC, Accuracy and F1 scores of ensemble models.

| Pair-wise models compared | AUC | | Accuracy | | F1 score | |
|---|---|---|---|---|---|---|
| | Mean difference (%) | p-value | Mean difference (%) | p-value | Mean difference (%) | p-value |
| CatBoost-AdaBoost | 04.98 | 0.00 | 05.85 | 0.00 | 06.63 | 0.00 |
| DT based Bagging-AdaBoost | 04.84 | 0.00 | 06.47 | 0.00 | 04.46 | 0.00 |
| GradientBoost-AdaBoost | 05.93 | 0.00 | 05.89 | 0.00 | 05.82 | 0.00 |
| Linear SVM-AdaBoost | 00.48 | 1.00 | 03.65 | 0.12 | 06.23 | 0.00 |
| RandomForest-AdaBoost | 04.13 | 0.01 | 04.24 | 0.04 | 03.70 | 0.03 |
| XGBoost-AdaBoost | 04.17 | 0.01 | 04.57 | 0.02 | 04.44 | 0.00 |
| DT-based Bagging-CatBoost | 09.82 | 0.00 | 12.32 | 0.00 | 11.09 | 0.00 |
| **GradientBoost-CatBoost *** | **00.95** | **0.99** | **00.04** | **1.00** | **00.81** | **0.99** |
| Linear SVM-CatBoost | 04.50 | 0.00 | 09.50 | 0.00 | 00.40 | 1.00 |
| **RandomForest-CatBoost *** | **00.85** | **0.99** | **01.62** | **0.91** | **02.93** | **0.17** |
| **XGBoost-CatBoost *** | **00.81** | **0.99** | **01.28** | **0.97** | **02.19** | **0.52** |
| GradientBoost-DT based Bagging | 10.77 | 0.00 | 12.36 | 0.00 | 10.28 | 0.00 |
| Linear SVM-DT based Bagging | 05.31 | 0.00 | 02.82 | 0.39 | 10.68 | 0.00 |
| RandomForest-DT based Bagging | 08.96 | 0.00 | 10.71 | 0.00 | 08.16 | 0.00 |
| XGBoost-DT based Bagging | 09.01 | 0.00 | 11.04 | 0.00 | 08.90 | 0.00 |
| Linear SVM-GradientBoost | 05.45 | 0.00 | 09.54 | 0.00 | 00.40 | 1.00 |
| **RandomForest-GradientBoost *** | **01.80** | **0.76** | **01.65** | **0.89** | **02.12** | **0.56** |
| **XGBoost-GradientBoost *** | **01.76** | **0.78** | **01.32** | **0.96** | **01.38** | **0.91** |
| RandomForest-Linear SVM | 03.65 | 0.05 | 07.88 | 0.00 | 02.53 | 0.34 |
| XGBoost-Linear SVM | 03.70 | 0.04 | 08.22 | 0.00 | 01.79 | 0.74 |
| **XGBoost-RandomForest *** | **00.05** | **1.00** | **00.34** | **1.00** | **00.74** | **1.00** |

***highlighted pairs are not significantly different**

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 May 2023 doi:10.20944/preprints202305.0426.v1



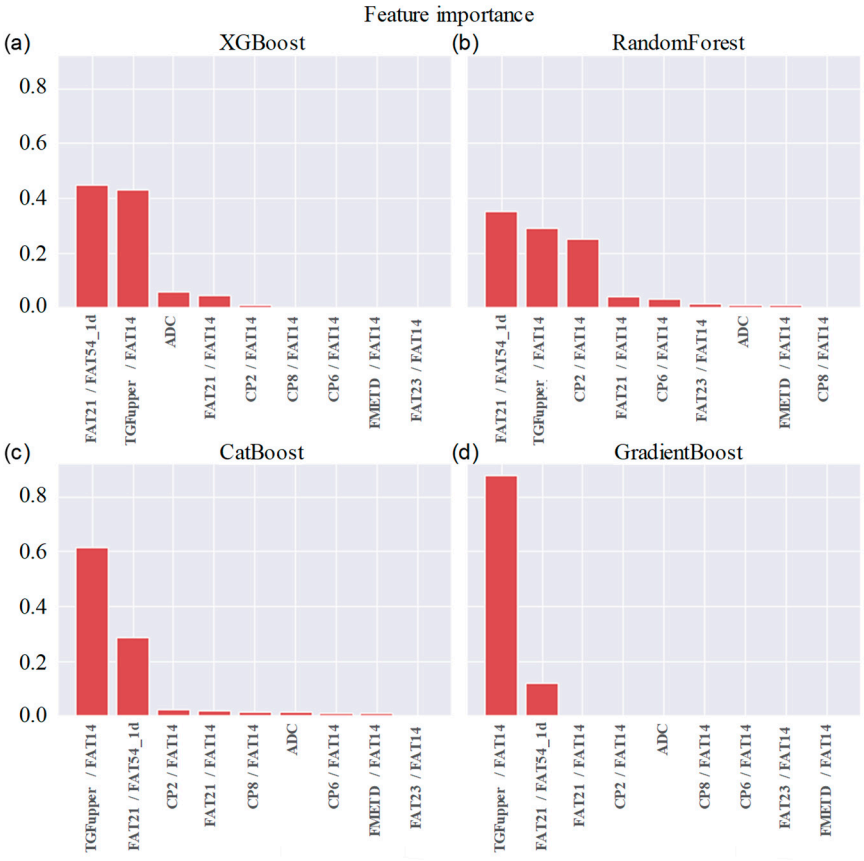
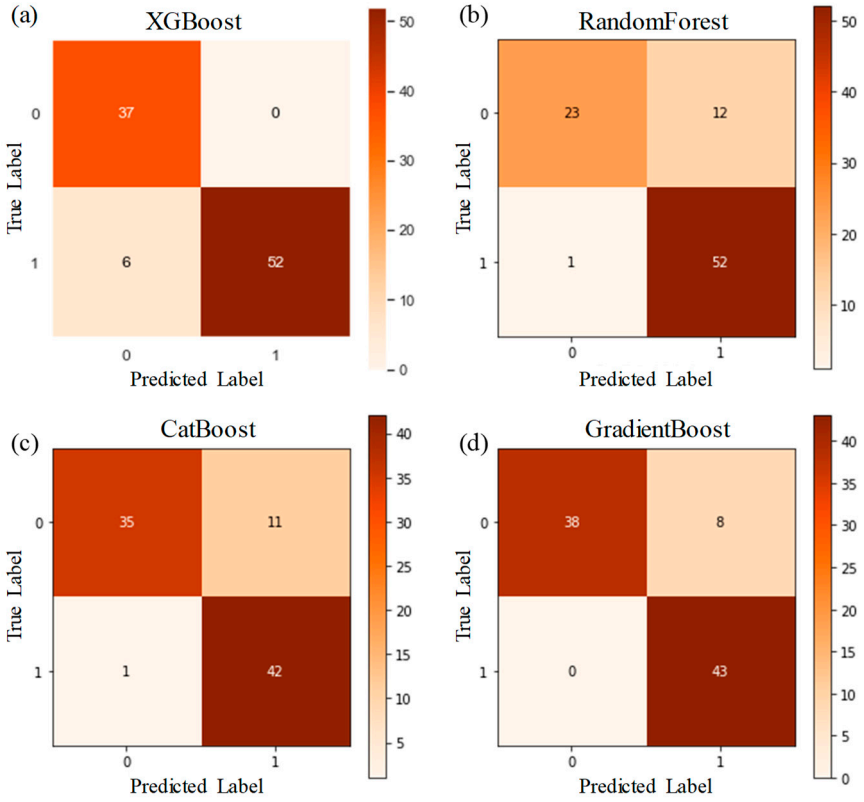



**Figure 6.** Bar charts showing the feature importance in (a) XGBoost, (b) RandomForest, (c) CatBoost and (d) GradientBoost.



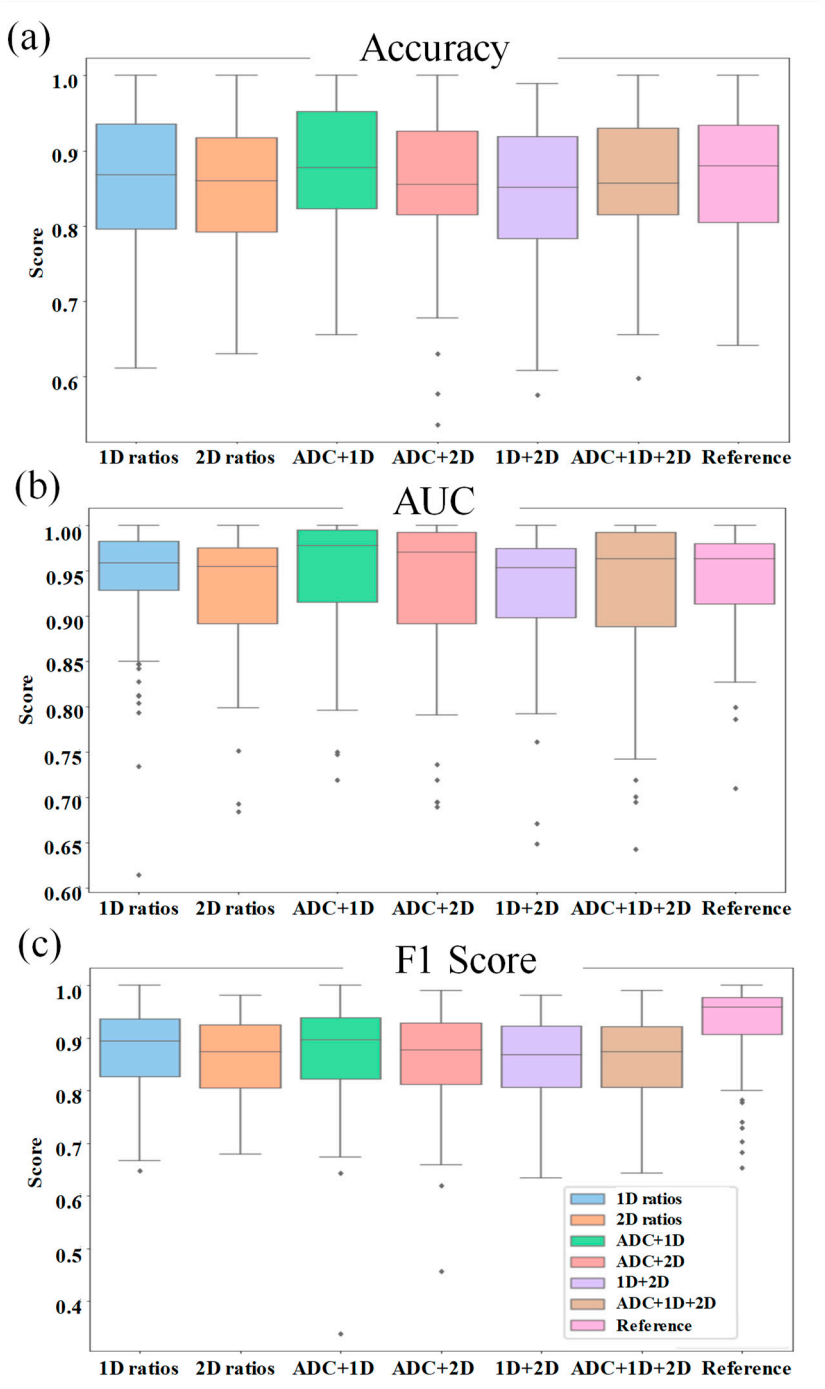
Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 May 2023 doi:10.20944/preprints202305.0426.v1



**Figure 7.** Confusion matrix of (a) XGBoost, (b) RandomForest, (c) CatBoost and (d) GradientBoost. Label 0 stands for benign and label 1 stands for malignant.

### *3.3. Ablation Study*

An ablation study was performed with choline, glycine + myo-Inositol, unsaturation index, and the features from non-water suppressed 1D spectra including water fraction, fat fraction and water-fat ratios. The model performance was also compared with and without ADC in the feature list. While the changes observed in the mean accuracy score over 100 repetitions were less than 1%, the change was positive in the presence of ADC (0.02%), while it was marginally negative in the presence of the additional metabolite ratios (<-0.009%). Boxplots comparing the performance matrices of the ablation study are shown in Figure 8.



 doi:10.20944/preprints202305.0426.v1



**Figure 8.** The box plots comparing the (a) accuracy, (b) F1 Score and (c) AUC metrics of the ensemble models during the ablation study. Performance of the models using the original nine features are shown on the right side in each panel as reference.

## 4. Discussion

This study showed the feasibility of using metabolite ratios from 5D EP-COSI and ADC values from the DWI data of breast cancer patients to train machine learning models for classifying benign and malignant lesions. While earlier studies have attempted to use lesion characterization using features extracted from the DWI and DCE-MRI data, these models did not use the quantitative measures of metabolite and lipid features which can be obtained with an MRSI examination (48-50). Although variations in water and fat levels can become ambiguous in glandular regions, especially in benign and healthy tissues, various lipid and metabolite ratios are reported to have statistically significant difference between the benign and malignant lesions (42). Building on this fact, our study pursued a detailed analysis of lesion characterization using 5D EP-COSI features in a machine learning framework.

The ensemble models were found to be performing better than the individual models. This is as expected, since they combine the strengths of multiple individual models (45). In fact, the ensemble models can use multiple base models to learn different aspects of the data and hence learn more complex relationships between the variables. They are also more robust to outliers and are also expected to reduce overfitting since they can compensate for the prediction errors of individual models. XGBoost, GradientBoost, RandomForest and CatBoost were found to be the best performing ensemble models in this study with 92% to 95% AUC, 85% to 87% accuracy and 87% to 90% F1 scores.

While the feature importance scores slightly varied among the top performing models, ADC was ranked among the top six. Four out of the top nine features were the ratios of cross-peaks, which are specific to the 2D COSY technique. The remaining four main features were the ratios of diagonal lipid peaks. It is interesting to note that the ratios of lipid cross peaks ranked higher than some of the conventional biomarkers like Cho and mI+Gly ratios for classifying benign and malignant lesions in the ML framework, despite both Cho and mI+Gly ratios being in the list of statistically significant variables in the MWU tests. The RFE process to select the significant features used XGBoost as the base model since it was one of the best-performing models based on all the variables from MWU tests. Therefore, it is to be noted that the top features were determined by their contribution in reducing loss function in the decision trees.

The number of datasets is one of the limitations in this study. Even though we have multiple voxels from the same dataset giving metabolite and lipid ratios, it is important to split the data based on the actual number of subjects rather than the voxels. It would be tempting to consider the individual voxels as separate data when splitting the training and testing sets. However, this approach could lead to severe data leakage, since multiple voxels from the same subject can have similar statistics, especially when interpolation is used to increase the number of voxels. Even otherwise, if the lesion spans multiple voxels in the spectroscopic data, the relatively low resolution and partial volume effects can potentially cause slightly overlapping information between the neighboring voxels. Therefore, if the train-test split is performed based on the voxels rather than individual subjects, it is reasonable to assume that during the training stage, the model would already see some of the statistics present in the testing data. This will artificially increase the score of test and validation performance matrices, but will not be generalizable to a new subject.

Even though these ML models should be generalizable to the MRSI/DWI data from different scanners and sites, it may be considered as another limitation of this study since there could be subtle/complex variations in the datasets from different scanners and sites so that the list of most important features could differ. A future study with a larger sample size ideally from different scanners and sites can further validate the results presented in this work.

Since the focus of this study was to analyze the performance of ML models with features from the 5D EP-COSI data, we have not considered some of the image-based features potentially available from DWI. For example, it has been recently shown that the features based on continuous-time

random-walk (CTRW) and intravoxel incoherent motion (IVIM) models from DWI using multiple b-values can classify benign and malignant breast lesions using ensemble ML models (48). The ablation study shows that the presence of ADC values in the list of features improves the score of performance matrices. Therefore, more features from DWI as well as other modalities like DCE-MRI can be used in a future study to potentially further improve the model performance.

## 5. Conclusion

In this pilot validation of the multi-dimensional (5D EP-COSI) data for characterization of breast tissues, we have shown that ML based classification models can be trained using spectroscopic features in conjunction with ADC values from DWI to classify benign and malignant lesions. Multiple diagonal and cross-peaks from 2D COSY spectra were identified as important features, further asserting the advantage of 2D COSY spectra as compared to features derived from 1D spectra. GradientBoost, CatBoost, RandomForest and XGBoost were the best performing models with 92% - 96% AUC, 85% - 87% Accuracy and 87% - 90% F1-scores.

**Author Contributions:** A.J. reconstructed the raw data, performed machine learning analysis, and prepared original draft; M.L. did the machine learning analysis; A.S. performed the data acquisition; M.J. and S.L.F. assisted with lesion analysis and manuscript editing; P.M.M. and A.S. edited the manuscript; A.C., U.E., J.B. proof-read the final draft; M.A.T. conceived the experimental design, acquired funding and edited the manuscript. All authors have read and agreed to the published version of the manuscript.

**Acknowledgments:** Authors like to acknowledge the scientific support of Dr. Manoj Sarma, Dr Zohaib Iqbal, Dr Brian Burns, Dr Neil Wilson, Ms. Kavya Umachandran and Ms. Samantha Joseph. Also, authors would like to thank the support of Ms. Victoria Rueda with the recruitment of study subjects and the UCLA Radiology MRI technicians during data collection.

**Funding:** This work was supported by a CDMRP grant from the US Army Breast Cancer Research Program:# W81XWH-16-1-0524.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Siegel RL, Miller KD, Wagle NS, Jemal A. Cancer statistics, 2023. CA Cancer J. Clin. 2023;73(1):17-48.
2. Bray F, Ferlay J, Soerjomataram I, Siegel R, Torre L, Jemal A. Erratum: Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J. Clin. 2020;70(4):313.
3. Al-Ajmi K, Lophatananon A, Yuille M, Ollier W, Muir KR. Review of non-clinical risk models to aid prevention of breast cancer. Cancer Causes Control. 2018;29(10):967-86.
4. Fu B, Liu P, Lin J, Deng L, Hu K, Zheng H. Predicting invasive disease-free survival for early stage breast cancer patients using follow-up clinical data. IEEE Trans. Biomed. Eng. 2018;66(7):2053-64.
5. Jagannathan N. Breast MR. NMR Biomed. 2009;22(1):1-2.
6. Lehman CD, Isaacs C, Schnall MD, Pisano ED, Ascher SM, Weatherall PT, et al. Cancer yield of mammography, MR, and US in high-risk women: prospective multi-institution breast cancer screening study. Radiology. 2007;244(2):381-8.
7. Morris EA. Diagnostic breast MR imaging: current status and future directions. Radiol. Clin. North Am. 2007;45(5):863-80.
8. Pe M, Dorme L, Coens C, Basch E, Calvert M, Campbell A, et al. Statistical analysis of patient-reported outcome data in randomised controlled trials of locally advanced and metastatic breast cancer: a systematic review. The Lancet Oncology. 2018;19(9):e459-e69.
9. Pop CF, Stanciu-Pop C, Drisis S, Radermeker M, Vandemerckt C, Noterman D, et al. The impact of breast MRI workup on tumor size assessment and surgical planning in patients with early breast cancer. The breast journal. 2018;24(6):927-33.
10. Saslow D, Boetes C, Burke W, Harms S, Leach MO, Lehman CD, et al. American Cancer Society guidelines for breast screening with MRI as an adjunct to mammography. CA Cancer J. Clin. 2007;57(2):75-89.
11. Weinreb JC, Newstead G. MR imaging of the breast. Radiology. 1995;196(3):593-610.
12. Zhang M, Horvat JV, Bernard-Davila B, Marino MA, Leithner D, Ochoa-Albiztegui RE, et al. Multiparametric MRI model with dynamic contrast-enhanced and diffusion-weighted imaging enables breast cancer diagnosis with high accuracy. J. Magn. Reson. Imaging. 2019;49(3):864-74.
13. Warner E. Breast-cancer screening. New England Journal of Medicine. 2011;365(11):1025-32.

14. Aboagye EO, Bhujwalla ZM. Malignant transformation alters membrane choline phospholipid metabolism of human mammary epithelial cells. Cancer Res. 1999;59(1):80-4.
15. Bolan PJ, Kim E, Herman BA, Newstead GM, Rosen MA, Schnall MD, et al. MR spectroscopy of breast cancer for assessing early treatment response: Results from the ACRIN 6657 MRS trial. J. Magn. Reson. Imaging. 2017;46(1):290-302.
16. Dorrius MD, Pijnappel RM, Jansen-van der Weide MC, Jansen L, Kappert P, Oudkerk M, et al. Determination of choline concentration in breast lesions: quantitative multivoxel proton MR spectroscopy as a promising noninvasive assessment tool to exclude benign lesions. New diagnostic developments to prevent unnecessary invasive procedures in breast cancer diagnostic work-up. 2011.
17. Gribbestad I, Sitter B, Lundgren S, Krane J, Axelson D. Metabolite composition in breast tumors examined by proton nuclear magnetic resonance spectroscopy. Anticancer Res. 1999;19(3A):1737-46.
18. Haukaas TH, Euceda LR, Giskeødegård GF, Bathen TF. Metabolic portraits of breast cancer by HR MAS MR spectroscopy of intact tissue samples. Metabolites. 2017;7(2):18.
19. Jagannathan N, Seenu V, Kumar M. Potential of in vivo proton MR spectroscopy in the assessment of breast lesions without the use of contrast agent. Radiology. 2002;223(1):281-2.
20. Roebuck JR, Cecil KM, Schnall MD, Lenkinski RE. Human breast lesions: characterization with proton MR spectroscopy. Radiology. 1998;209(1):269-75.
21. Sharma U, Mehta A, Seenu V, Jagannathan N. Biochemical characterization of metastatic lymph nodes of breast cancer patients by in vitro 1H magnetic resonance spectroscopy: a pilot study. Magn. Reson. Imaging. 2004;22(5):697-706.
22. Thakur SB, Horvat JV, Hancu I, Sutton OM, Bernard-Davila B, Weber M, et al. Quantitative in vivo proton MR spectroscopic assessment of lipid metabolism: Value for breast cancer diagnosis and prognosis. J. Magn. Reson. Imaging. 2019;50(1):239-49.
23. Thomas MA, Binesh N, Yue K, DeBruhl N. Volume-localized two-dimensional correlated magnetic resonance spectroscopy of human breast cancer. Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine. 2001;14(2):181-6.
24. Amornsiripanitch N, Nguyen VT, Rahbar H, Hippe DS, Gadi VK, Rendi MH, et al. Diffusion-weighted MRI characteristics associated with prognostic pathological factors and recurrence risk in invasive ER+/HER2–breast cancers. J. Magn. Reson. Imaging. 2018;48(1):226-36.
25. Bammer R. Basic principles of diffusion-weighted imaging. Eur. J. Radiol. 2003;45(3):169-84.
26. Belli P, Costantini M, Bufi E, Magistrelli A, La Torre G, Bonomo L. Diffusion-weighted imaging in breast lesion evaluation. Radiol. Med. 2010;115(1):51-69.
27. Delbany M, Bustin A, Poujol J, Thomassin-Naggara I, Felblinger J, Vuissoz PA, et al. One-millimeter isotropic breast diffusion-weighted imaging: Evaluation of a superresolution strategy in terms of signal-to-noise ratio, sharpness and apparent diffusion coefficient. Magn. Reson. Med. 2019;81(4):2588-99.
28. deSouza NM. Diffusion-weighted MRI in multicenter trials of breast cancer: a useful measure of tumor response? : Radiological Society of North America; 2018. p. 628-9.
29. Le Bihan D, Breton E, Lallemand D, Grenier P, Cabanis E, Laval-Jeantet M. MR imaging of intravoxel incoherent motions: application to diffusion and perfusion in neurologic disorders. Radiology. 1986;161(2):401-7.
30. Newitt DC, Zhang Z, Gibbs JE, Partridge SC, Chenevert TL, Rosen MA, et al. Test–retest repeatability and reproducibility of ADC measures by breast DWI: Results from the ACRIN 6698 trial. J. Magn. Reson. Imaging. 2019;49(6):1617-28.
31. Sharma U, Danishad KKA, Seenu V, Jagannathan NR. Longitudinal study of the assessment by MRI and diffusion-weighted imaging of tumor response in patients with locally advanced breast cancer undergoing neoadjuvant chemotherapy. NMR in Biomedicine: An International Journal Devoted to the Development and Application of Magnetic Resonance In vivo. 2009;22(1):104-13.
32. Furman-Haran E, Grobgeld D, Kelcz F, Degani H. Critical role of spatial resolution in dynamic contrast-enhanced breast MRI. Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine. 2001;13(6):862-7.
33. Hickman P, Moore N, Shepstone B. The indeterminate breast mass: assessment using contrast enhanced magnetic resonance imaging. The British Journal of Radiology. 1994;67(793):14-20.
34. Kvistad KA, Rydland J, Vainio J, Smethurst HB, Lundgren S, Fjøsne HE, et al. Breast lesions: evaluation with dynamic contrast-enhanced T1-weighted MR imaging and with T2*-weighted first-pass perfusion MR imaging. Radiology. 2000;216(2):545-53.
35. Liu P, Debatin J, Caduff R, Kacl G, Garzoli E, Krestin G. Improved diagnostic accuracy in dynamic contrast enhanced MRI of the breast by combined quantitative and qualitative analysis. The British journal of radiology. 1998;71(845):501-9.
36. Millet I, Pages E, Hoa D, Merigeaud S, Curros Doyon F, Prat X, et al. Pearls and pitfalls in breast MRI. The British journal of radiology. 2012;85(1011):197-207.

37. Bogner W, Gruber S, Pinker K, Grabner G, Stadlbauer A, Weber M, et al. Diffusion-weighted MR for differentiation of breast lesions at 3.0 T: how does selection of diffusion protocols affect diagnosis? Radiology. 2009;253(2):341-51.
38. Chen X, Li W-l, Zhang Y-l, Wu Q, Guo Y-m, Bai Z-l. Meta-analysis of quantitative diffusion-weighted MR imaging in the differential diagnosis of breast lesions. BMC Cancer. 2010;10(1):1-11.
39. Prvulovic Bunovic N, Sveljo O, Kozic D, Boban J. Is Elevated Choline on Magnetic Resonance Spectroscopy a Reliable Marker of Breast Lesion Malignancy? Front. Oncol. 2021;11:610354.
40. Shahraki Z, Ghaffari M, Parooie F, Salarzaei M. Preoperative evaluation of breast cancer: Contrast-enhanced mammography versus contrast-enhanced magnetic resonance imaging: A systematic review and meta-analysis. Breast Dis. 2022;41(1):303-15.
41. Baltzer PA, Dietzel M. Breast lesions: diagnosis by using proton MR spectroscopy at 1.5 and 3.0 T—systematic review and meta-analysis. Radiology. 2013;267(3):735-46.
42. Joy A, Saucedo A, Joines M, Lee-Felker S, Kumar S, Sarma MK, et al. Correlated MR spectroscopic imaging of breast cancer to investigate metabolites and lipids: acceleration and compressed sensing reconstruction. BJR| Open. 2022;4:20220009.
43. Aamir S, Rahim A, Aamir Z, Abbasi SF, Khan MS, Alhaisoni M, et al. Predicting breast cancer leveraging supervised machine learning techniques. Comput. Math. Methods Med. 2022;2022.
44. Dou Y, Meng W. An Optimization Algorithm for Computer-Aided Diagnosis of Breast Cancer Based on Support Vector Machine. Frontiers in Bioengineering and Biotechnology. 2021;9:698390.
45. Greener JG, Kandathil SM, Moffat L, Jones DT. A guide to machine learning for biologists. Nature Reviews Molecular Cell Biology. 2022;23(1):40-55.
46. Zhou Z-H, Zhou Z-H. Ensemble learning: Springer; 2021.
47. Qi C, Li Y, Fan X, Jiang Y, Wang R, Yang S, et al. A quantitative SVM approach potentially improves the accuracy of magnetic resonance spectroscopy in the preoperative evaluation of the grades of diffuse gliomas. NeuroImage: Clinical. 2019;23:101835.
48. Mehta R, Bu Y, Zhong Z, Dan G, Zhong P-S, Zhou C, et al. Characterization of breast lesions using multi-parametric diffusion MRI and machine learning. Phys. Med. Biol. 2023;68(8):085006.
49. Parekh VS, Jacobs MA. Integrated radiomic framework for breast cancer and tumor biology using advanced machine learning and multiparametric MRI. NPJ breast cancer. 2017;3(1):43.
50. Daimiel Naranjo I, Gibbs P, Reiner JS, Lo Gullo R, Sooknanan C, Thakur SB, et al. Radiomics and machine learning with multiparametric breast MRI for improved diagnostic accuracy in breast cancer diagnosis. Diagnostics. 2021;11(6):919.
51. Klose U. In vivo proton spectroscopy in presence of eddy currents. Magn. Reson. Med. 1990;14(1):26-30.
52. Burns BL, Wilson NE, Thomas MA. Group sparse reconstruction of multi-dimensional spectroscopic imaging in human brain in vivo. Algorithms. 2014;7(3):276-94.
53. Wilson NE, Burns BL, Iqbal Z, Thomas MA. Correlated spectroscopic imaging of calf muscle in three spatial dimensions using group sparse reconstruction of undersampled single and multichannel data. Magn. Reson. Med. 2015;74(5):1199-208.
54. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Machine learning. 2002;46:389-422.
55. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. the Journal of machine Learning research. 2011;12:2825-30.
56. Buitinck L, Louppe G, Blondel M, Pedregosa F, Mueller A, Grisel O, et al. API design for machine learning software: experiences from the scikit-learn project. arXiv preprint arXiv:1309.0238. 2013.
57. Chen T, Guestrin C, editors. Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining; 2016.
58. Dorogush AV, Ershov V, Gulin A. CatBoost: gradient boosting with categorical features support. arXiv preprint arXiv:1810.11363. 2018.