electronics

MDPI

Article

Mammographic Classification of Breast Cancer Microcalcifications through Extreme Gradient Boosting

Haobang Liang [1,†], Jiao Li [2,†], Hejun Wu [3,*], Li Li [2,*], Xinrui Zhou [3] and Xinhua Jiang [2]

[1] School of Biomedical Engineering, Sun Yat-sen University, Guangzhou 510006, China
[2] Radiology Department, Sun Yat-sen University Cancer Center, Guangzhou 510060, China
[3] Department of Computer Science, Sun Yat-sen University, Guangzhou 510006, China
* Correspondence: wuhejun@mail.sysu.edu.cn (H.W.); li2@mail.sysu.edu.cn (L.L.)
† First authors: Haobang Liang and Jiao Li.

Abstract: In this paper, we proposed an effective and efficient approach to the classification of breast cancer microcalcifications and evaluated the mathematical model for calcification on mammography with a large medical dataset. We employed several semi-automatic segmentation algorithms to extract 51 calcification features from mammograms, including morphologic and textural features. We adopted extreme gradient boosting (XGBoost) to classify microcalcifications. Then, we compared other machine learning techniques, including k-nearest neighbor (kNN), adaboostM1, decision tree, random decision forest (RDF), and gradient boosting decision tree (GBDT), with XGBoost. XGBoost showed the highest accuracy (90.24%) for classifying microcalcifications, and kNN demonstrated the lowest accuracy. This result demonstrates that it is essential for the classification of microcalcification to use the feature engineering method for the selection of the best composition of features. One of the contributions of this study is to present the best composition of features for efficient classification of breast cancers. This paper finds a way to select the best discriminative features as a collection to improve the accuracy. This study showed the highest accuracy (90.24%) for classifying microcalcifications with AUC = 0.89. Moreover, we highlighted the performance of various features from the dataset and found ideal parameters for classifying microcalcifications. Furthermore, we found that the XGBoost model is suitable both in theory and practice for the classification of calcifications on mammography.

Keywords: adaboostM1; breast cancer; classification; GBDT; kNN; mammographic; microcalcifications; RDF; XGBoost

Citation: Liang, H.; Li, J.; Wu, H.; Li, L.; Zhou, X.; Jiang, X. Mammographic Classification of Breast Cancer Microcalcifications through Extreme Gradient Boosting. *Electronics* **2022**, *11*, 2435. https://doi.org/10.3390/electronics11152435

Academic Editor: Rui Pedro Lopes

Received: 30 June 2022
Accepted: 2 August 2022
Published: 4 August 2022

1. Introduction

Breast cancer is a common but critical disease of women in the world, as stated in the World Health Organization GLOBOCAN 2012 report [1]. In China, breast cancer is currently one of the most common causes of death. Due to China's large population, approximately 11% of worldwide breast cancers occur in China [2]. Moreover, breast cancer patients in China tend to be younger and have denser breasts. To date, various conventional methods, such as infrared, X-ray mammography, computed tomography, ultrasound and magnetic resonance imaging, have been widely used for breast tumor diagnosis [3]. Mammography is one of the most reliable methods for early detection as well as reduction of mortality [4–6]. Given the enormous size of the screened population, interpreting mammograms, even by experienced radiologists, can be both time and energy-consuming. Computer-aided detection and diagnosis (CAD) systems have long been studied as alternatives to save time and minimize subjectivity. The Breast Imaging Reporting and Data System (BI-RADS) lexicon, by American College of Radiology, provides standard mammographic reports to facilitate biopsy decision-making [7]. Both the sensitivity and efficiency of mass detection for mammography are still low. In mammography, clustered microcalcifications are the

main warning signs for cancer and sometimes may be the only signs. Typical malignant or benign microcalcifications can be classified on the basis of their distribution and morphologic features. In addition, studies have shown that 90% of non-palpable in situ ductal carcinomas and 70% of non-palpable minimal carcinomas were visible as microcalcifications alone during the screening process. Therefore, accurate as well as rapid identification of malignant calcifications improves the true positive detection rate. Traditionally, CAD in mammography relies on pattern recognition and classification. Classical image processing and machine learning techniques were combined to identify suspicious calcifications and to differentiate among types. Traditional CAD systems perform image segmentation, feature extraction for calcification, as well as classification [8]. For breast cancer, the selection of the classification method is critical. Several algorithms can be applied to the effort, including k-nearest neighbor (kNN), adaboostM1, decision tree, random decision forest (RDF), and gradient boosting decision tree (GBDT). Indeed, these techniques have already been developed for and applied in breast cancer research. Among these techniques, one of the most widely used approaches is tree boosting. Recently, in academia and the intelligence industry, a scalable end-to-end tree boosting system called extreme gradient boosting (XGBoost) has been employed in a number of machine learning and data mining challenges and has been successfully applied to many classification problems to achieve what are considered as state-of-art results [9]. XGBoost learns an ensemble of decision trees. This boosting technique was adjusted to enhance a Taylor expansion of the loss functions. In this study, we employed extreme gradient boosting to discriminate among microcalcifications in mammograms automatically. Moreover, we analyze and select the best composition of features from various extracted features provided by a CAD system related to breast cancer.

2. Related Work

This paper presents the classification of breast cancer based on a computer-aided diagnosis (CAD) system. A standard CAD system is capable of segmenting structures, detecting abnormalities, and extracting features. Algorithms have been proposed to evaluate the classified features extracted from CAD systems. These algorithms for CAD systems must consider sensitivity, specificity, and evaluation of positive predictions. For instance, algorithms can discriminate various stages of cancer using texture characteristics with these algorithms [10,11]. Feature selection is often used to reduce the dimension of data in order to improve the efficiency of data processing [12,13]. In comparison with previous studies, this paper is able to select the best composition of features, through coordinate descent.

In recent years, the radiological evaluation of breast cancer has focused on microcalcification or masses, with relatively more attention on microcalcification. A wide range of machine learning algorithms have been developed for the early diagnosis of breast cancers [14–16]. The most common algorithms are based on the algorithm of k-nearest neighbor (kNN) [17,18], adaboostM1, and a series of tree models, such as decision tree, random decision forest (RDF), and gradient boosting decision tree (GBDT).

For a given dataset with n examples and m features $D = (x_i, y_i)(|D| = n, x_i \in R^n, y_i \in \{0, 1\})$, a tree ensemble model uses K additive functions to predict the output. The final score, shown in Equation (1), also represents the differentiable loss function that measures the difference between the predicted $\hat{y}_i$ and the target y_i:

$$\hat{y}_i = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i), f_k \in \mathcal{F} \quad (1)$$

where f_k and $\mathcal{F}$ represent the independent tree structure with leaf scores and the space of all regression trees (also known as CART), respectively. Equation (2) is used to optimize the regularized objective:

$$L(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \quad (2)$$

where $\Omega(f) = \gamma T + \frac{1}{2}\lambda \parallel \omega \parallel^2$.

$\Omega(f)$ is present to avoid overfitting, which penalizes the complexity of the model. It is given by Equation (3), where T represents the number of leaves and ω represents the weight of each leaf. λ and γ are two constants to control the regularization degree.

$$\gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{3}$$

From the use of regularization, two additional techniques are used to prevent overfitting further [7]. The first technique is shrinkage [19], and the second is column (feature) subsampling, also called random forest [20,21].

However, Equation (2) includes functions as parameters that cannot be optimized using traditional optimization methods. Let $y_i^{(t)}$ be the prediction of the i-th instance at the t-th iteration. We will need to add ft to minimize the following objective [9], thus Equation (2) can be written as Equation (4).

$$L^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) \tag{4}$$

Furthermore, a second-order approximation can be used to quickly optimize the objective in the general setting [22] as shown in Equation (5):

$$L^{(t)} \simeq \sum_{i=1}^{n} \left[l\left(y_i, \hat{y}^{(t-1)}\right) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) \tag{5}$$

where $g_i = \partial_{\hat{y}^{(t-1)}} l\left(y_i, \hat{y}^{(t-1)}\right)$ and $h_i = \partial^2_{\hat{y}^{(t-1)}} l\left(y_i, \hat{y}^{(t-1)}\right)$ are first and second-order gradient statistics on the loss function, respectively. Moreover, a simplified objective function without constants at step t is shown in Equation (6):

$$\widetilde{L}^{(t)} = \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) \tag{6}$$

Equation (7) is an objective function generated by expanding the regularization term:

$$\begin{aligned} \widetilde{L}^{(t)} &= \sum_{i=1}^{n} \left[g_i f_t(x_i) + \tfrac{1}{2} h_i f_t^2(x_i) \right] + \gamma T + \tfrac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \\ &= \sum_{j=1}^{T} \left[(\sum_{i \in I_j} g_i) w_j + \tfrac{1}{2} (\sum_{i \in I_j} h_i + \lambda) w_j^2 \right] + \gamma T \end{aligned} \tag{7}$$

Here, we define equation $I_j = \{q(x_i) = j\}$ as the instance set of the leaf j. For a fixed structure $q_{(x)}$, we can compute the optimal weight w_j^* of the leaf j by Equation (8), and calculate the corresponding optimal value by Equation (9). It can be used as a scoring function to measure the quality of a tree structure q.

$$w_j^* = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda} \tag{8}$$

$$\widetilde{L}^{(t)}(q) = -\frac{1}{2} \sum_{j=1}^{T} \frac{\left(\sum_{i \in I_j} g_i\right)^2}{\sum_{i \in I_j} h_i + \lambda} + \gamma T \tag{9}$$

Equation (10) is the loss reduction after the split, usually used in practice for evaluating split candidates. Letting $I = I_L \cup I_R$, thus I_L and I_R are the instance sets of left and right nodes after the split, and $G_j = I \in g_I$ and $H_j = I \in h_i$.

$$L_{split} = \frac{1}{2} \left[\frac{\left(\sum_{i \in I_L} g_i\right)^2}{\sum_{i \in I_L} h_i + \lambda} + \frac{\left(\sum_{i \in I_R} g_i\right)^2}{\sum_{i \in I_R} h_i + \lambda} - \frac{\left(\sum_{i \in I} g_i\right)^2}{\sum_{i \in I} h_i + \lambda} \right] - \gamma \tag{10}$$

3. Image Segmentation and Feature Selection

Images were obtained on a GE Senographe DS mammography system and a Siemens Mammomat Inspiration mammography system. All images were digitized at 1024 × 1024 pixels and an 8-bit greyscale level. Data regarding microcalcifications were extracted through image segmentation. Both statistical and textural features were used to classify image features. To obtain a comprehensive characterization of microcalcifications, we considered various types of features that have been widely used in studies of breast lesions as input data rather than original images [23–25]. A total of 13 features were extracted from regions of interest (ROI) and were recorded for each patient. These extracted features were intended to provide as comprehensive a characterization of the image as possible. These consisted of intensity, statistic, shape, and texture features. These features have been extensively tested in breast lesion studies [23–27].

The features are not selected randomly. The main warning signs and even the only signs for breast cancer alone have low sensitivity and efficiency. In this paper, we aim to find the best composition of features for efficient and accurate classification of breast cancers. Our method selects a feature to be included in the feature set using a process similar to the coordinate ascent. Initially, the feature set only includes the most discriminative feature. Then, features are selected one by one, given the fact that adding a feature can significantly improve the performance of classification. This process continues until adding a feature cannot improve the performance or can even deteriorate the performance. This way, the method is able to select the best discriminative features as a collection to improve the accuracy.

All features were selected to represent various dimensional aspects of microcalcifications, including one-dimensional shape features (average diameter), two-dimensional morphological features (area), dimensional fractal features (density, circularity proportion, solidity, sandy microcalcification, spiculation, and volume ratio), grey-level intensity statistics features (mean grey value), and statistics features (microcalcification number, circularity, and linear microcalcification). To increase the diversity of features and optimize experimental conditions, we selected 38 features with the approach of texture features in MATLAB. Two popular methods estimated the texture features, grey-level co-occurrence matrix (GLCM) [28–30], and grey-level run length matrix (GLRLM) [31,32]. The GLCM was calculated by counting the number of times adjacent pixels have the same orientation. These features can characterize the scattering of calcification satisfactorily.

The definition of these features (autocorrelation, contrast, cluster prominence, cluster shade, dissimilarity, energy, entropy, homogeneity, maximum probability, the sum of squares, sum average, sum variance, sum entropy, difference variance, difference entropy, information measure of correlation, inverse difference normalized, and inverse difference moment normalized) can be found in the MATLAB toolbox.

4. Results

The training group consisted of 5476 images, including 2813 benign and 2663 malignant lesions. All parameters were estimated by 10-fold cross-validation on the training group. Data regarding microcalcifications were extracted through image segmentation in CAD. All histopathological features, including statistical and textural features, were used to classify image features and obtain a comprehensive characterization of microcalcifications.

We recorded 51 features for each patient, including 38 features made from grey covariance matrix in two dimensions. All features were fed into kNN, adaboostM1, decision tree, RDF, GBDT, and XGBoost algorithms. If there are N samples and each of which has dimension D, the complexity of kNN is $O(N \times D)$. The complexity of adaboostM1 is $O(N \times D^2)$. The complexity of decision tree is $\mathrm{O}\left(\mathrm{N}^2 \times \mathrm{D} \times \log(\mathrm{D})\right)$. The complexity of RDF is $O(M(DNlogN))$, where M is the number of trees. The complexity of XGBoost is $O(DNlogN) + O(KDNE)$, where K is the number of trees and E is the depth of trees. To normalize the images and improve processing efficiency, we extracted the region of interest (ROI) first by a coarse segmentation scheme. The coarse segmentation procedure used

Otsu's method and morphological filters. Then, the resulting image was dilated using a dilation filter to obtain maximal connected region as calcification area.

The segmentation scheme is illustrated in Figure 1. Figure 2 displays an example of the automatic detection and segmentation pipeline for suspicious microcalcifications in the left breast of a 56-year-old patient with invasive ductal carcinoma. The microcalcifications were extracted from the raw data to delineate the image characteristics (Figure 2b).

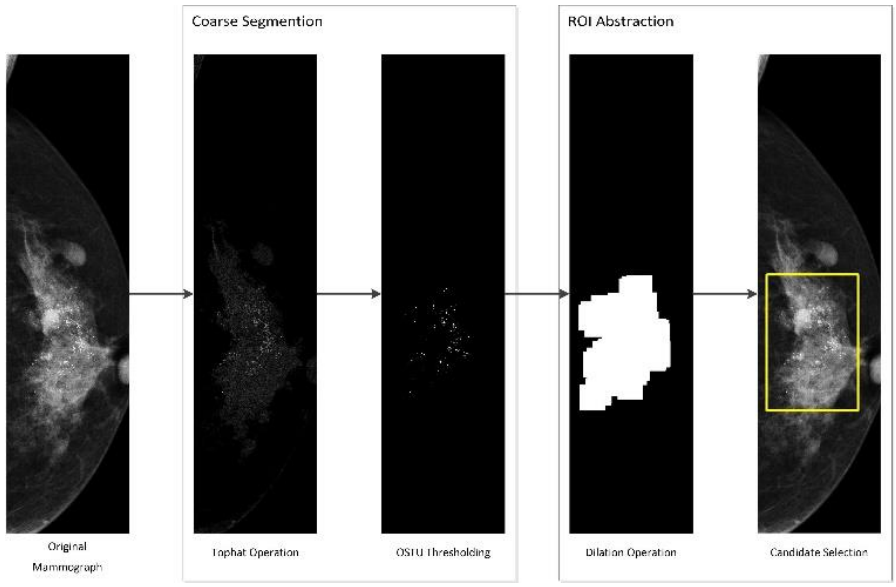

Figure 1. Workflow diagram for ROI abstraction.

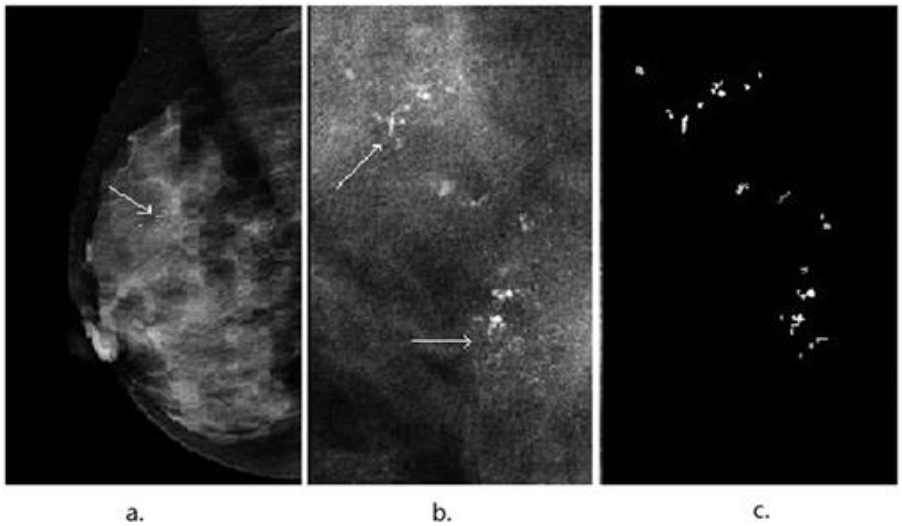

Figure 2. An illustrative example showing segmentation of microcalcifications in a mammogram of the left breast of a 56-year-old patient. (**a**) The mediolateral oblique (MLO) view shows clustered coarse and low-density microcalcifications (indicated by thin arrows). (**b**) The image shows the region of suspicious microcalcifications (indicated by thin arrows). (**c**) The segmented microcalcifications from (**b**) are used to characterize the features.

Figure 3 shows the image of the right breast of a 49-year-old patient with fibrocystic changes in which the focal microcalcifications appear as low contrast compared with the high-density background. Then, we combined texture features and the actual situation and obtained 51 features through the image.

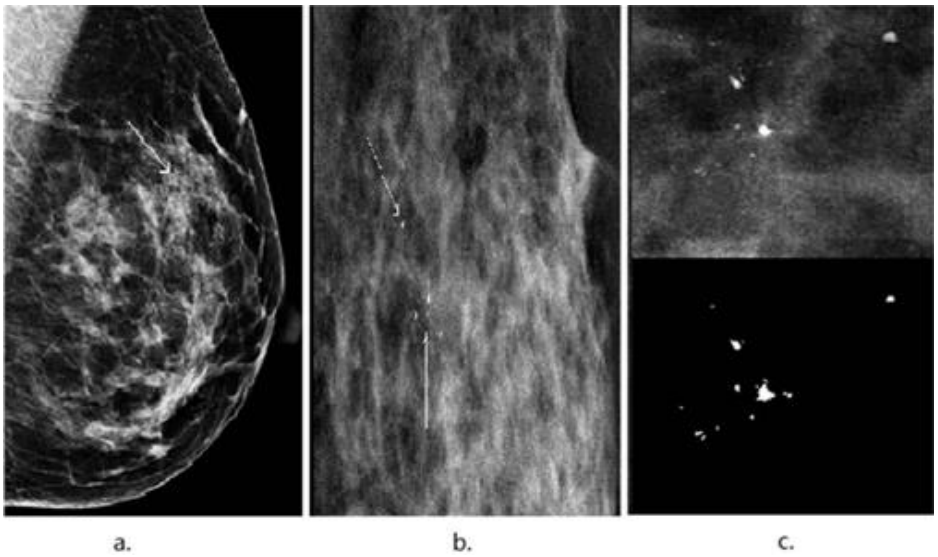

Figure 3. An illustrative example showing segmentation of microcalcifications in a mammogram of the right breast of a 49-year-old patient with fibrocystic changes. (**a**) The focal microcalcifications (indicated by thin arrows) appear as low contrast compared with the dense background in the mediolateral oblique (MLO) view. (**b**) Thin arrows indicate the region of suspicious microcalcifications. (**c**) A zoomed-in view of (**b**) highlights the segmented microcalcifications.

To evaluate the performance and discriminative power of every technique, we made quantitative measurements for overall classification accuracy, precision, recall, and F1-score, as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{11}$$

$$Recall = \frac{TP}{TP + TN} \tag{12}$$

$$Precision = \frac{TP}{TP + FP} \tag{13}$$

$$F1 = \frac{2TP}{2TP + FN + FP} \tag{14}$$

ROC indicates the receiver operating characteristic, which is a graphical plot that illustrates the diagnostic ability of classifier systems as their discrimination threshold is varied. The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. AUC indicates the area under the ROC curve.

Earlier experiments suggested that the classifiers' discriminative performance can be increased through the comprehensive characterization of microcalcifications rather than the characterization of individual features. Therefore, this approach was used in the following experiments.

Three scenarios for discriminating between benign and malignant lesions were experimented: The result of all the raw microcalcifications features; the result of raw combined with GLCM texture features; the result of feature selection via the random forest. The three scenarios' primary aims were to investigate the power of various features of microcalcifications and increase the number of features to improve the strength generalization ability of the model. Feature selection helps in reducing the influence of weak correlation features. The results were compared to those of kNN, adaboostM1, decision tree, RDF, GBDT, and XGBoost benchmark classifiers.

In the first scenario, image segmentation yielded 13 raw features. The overall accuracies were 64.0%, 84.8%, 85.1%, 85.1%, 85.5%, and 87.3% for kNN, decision tree, adaboostM1, RDF, GBDT, and XGBoost, respectively (Table 1).

Table 1. Overall result performance of all algorithms (13 features).

Algorithms	Accuracy	Recall	Precision	F1-Score	AUC
kNN	63.99%	0.6612	0.6263	0.6433	0.6407
Decision Tree	84.78%	0.8511	0.8390	0.8511	0.8501
adaboostM1	85.07%	0.8734	0.8451	0.8590	0.8504
random forest	85.08%	0.8526	0.8514	0.8378	0.8501
GBDT	85.45%	0.8801	0.8595	0.8696	0.8695
XGBoost	87.21%	0.8750	0.8597	0.8697	0.8699

In the second scenario, the image segmentation process yielded 51 features; all the experimental data results generally increased by two percentage points than the first scenario, and achieved the highest results among the three scenarios. The overall accuracies were 65.1%, 86.9%, 85.3%, 87.3%, and 88.7%, 90.2% for kNN, decision tree, adaboostM1, RDF, GBDT, and XGBoost, respectively (Table 2). XGBoost achieved the highest accuracy and AUC values (90.24% and 0.8903, respectively).

Table 2. Overall result performance of all algorithms (51 features).

Algorithms	Accuracy	Recall	Precision	F1-Score	AUC
kNN	65.06%	0.6673	0.6350	0.6500	0.6707
Decision Tree	86.89%	0.8701	0.8514	0.8378	0.8520
adaboostM1	85.27%	0.8734	0.8300	0.8590	0.8527
random forest	87.29%	0.8774	0.8672	0.8643	0.8629
GBDT	88.74%	0.8801	0.8765	0.8758	0.8774
XGBoost	90.24%	0.8845	0.9000	0.8952	0.8903

In the third scenario, based on the 51 features, we obtained the top 15 features (Table 3) ranked by random forest. The overall accuracies were somewhat lower than those of the second scenario: 65.3%, 85.2%, 85.8%, 86.3%, 89.2% for kNN, adaboostM1, RDF, GBDT, and XGBoost, respectively. Furthermore, the performance of GBDT was only marginally higher than the adaboostM1 and RDF model, while the accuracy of XGBoost exceeded GBDT by about 3% (Table 4).

Table 3. Top 15 important calcification features after feature selection.

Rank	Feature	Remark
1	number of calcification spots	morphologic features
2	percentage of gravel calcification	morphologic features
3	sum average	texture features
4	sum entropy	texture features
5	average diameter of calcification	morphologic features
6	percentage of circular degree	morphologic features
7	number of linear calcification point	morphologic features
8	circular degree	morphologic features
9	axis ratio	morphologic features
10	proportion of calcification	morphologic features
11	entity	morphologic features
12	volume rate	morphologic features
13	difference entropy	texture features
14	difference variance	texture features
15	average grey-level	morphologic features

Table 4. Overall result performance of all algorithms (15 features).

Algorithms	Accuracy	Recall	Precision	F1-Score	AUC
kNN	64.90%	0.6566	0.6059	0.6303	0.6408
Decision Tree	85.04%	0.8290	0.8490	0.8429	0.8411
adaboostM1	85.17%	0.8367	0.8300	0.8428	0.8426
random forest	85.77%	0.8249	0.8574	0.8456	0.8519
GBDT	85.90%	0.8670	0.8627	0.8598	0.8591
XGBoost	88.13%	0.8713	0.8810	0.8690	0.8792

These findings confirmed that, by accessing a large dataset, XGBoost produced the highest accuracy, showing an excellent capacity to discriminating between benign and malignant lesions through mammography, compared with standard models. Our model achieved similar outcomes in agreement with these reports, as demonstrated by the ROC curves in Figure 4.

These ROC curves compare the discriminative performances of individual features versus combinations of features. The accuracy of the XGBoost model exceeded 90%, and the kNN returned the worst performance, with nearly 63% accuracy. AdaboostM1, decision tree, and RDF gave similar results, with both higher than kNN. GBDT was slightly better than RDF, achieving the second-highest accuracy (88%) in both three scenarios.

To compare whether the prediction error rate of XGBoost and other models are significantly different, we used Kolmogorov-Smirnov (KS) predictive accuracy (KSPA) test [33] on different features sets. The KSPA test consists of a two-sided KS test followed by a one-sided KS test to check for model errors. The two-sided KS test checked significant statistical differences between the two models (when p-value is less than 0.05). The one-sided KS test conveys whether the model provides a smaller random error rate (also when p-value is less than 0.05).

In this paper, the absolute value error of each model was used as input in the KSPA test, and we defined the significance level as 0.05. The experimental results (Tables 5–7) indicate that there is indeed a significant statistical difference in the prediction errors between XGBoost and other models. Apart from RDF, XGBoost has lower prediction errors than other models in the prediction of these three features of quantity sets. Although there is no significant difference between XGBoost and RDF on 13 and 15 features sets, one-sided KS test provides sufficient evidence that XGBoost has a lower random error rate than other models on 51 features sets.

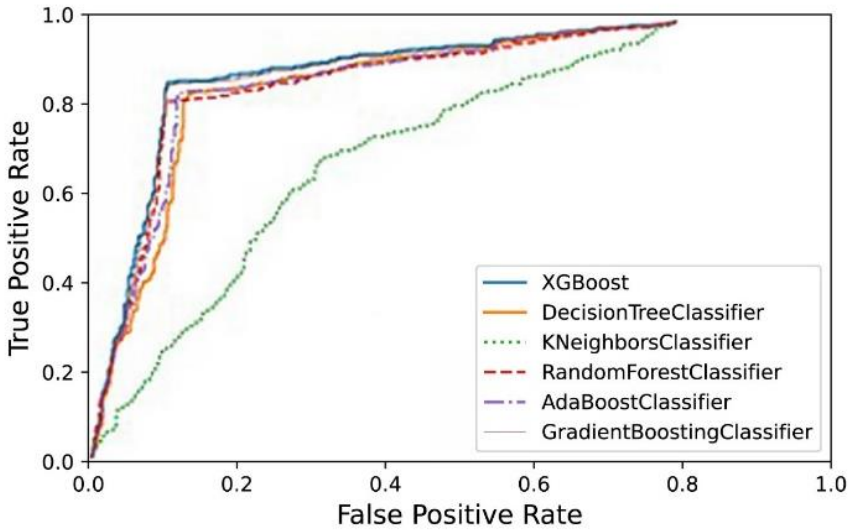

Figure 4. *Cont.*

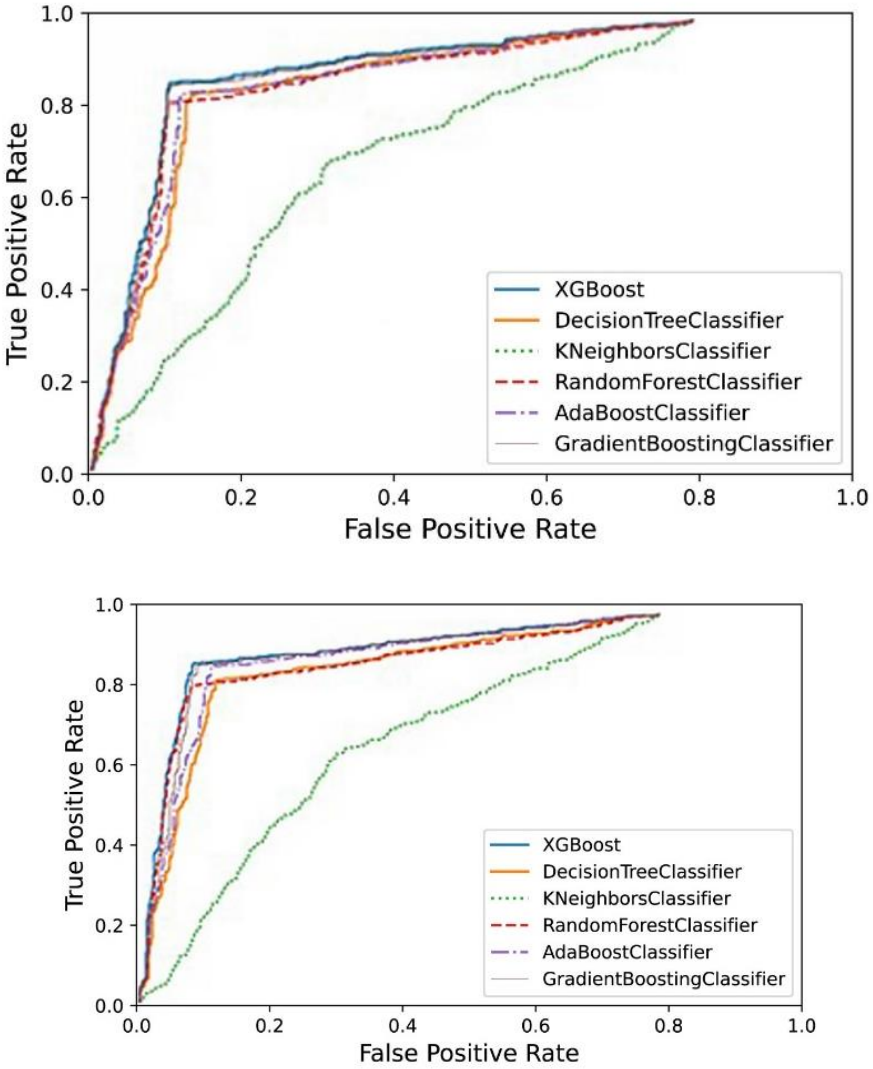

Figure 4. These ROC curves compare the discriminative performances of individual features versus combinations of features.

Table 5. Kolmogorov-Smirnov predictive accuracy test (13 features).

	kNN vs. XGBoost	**Decision Tree vs. XGBoost**	**adaboostM1 vs. XGBoost**	**Random Forest vs. XGBoost**	**GBDT vs. XGBoost**
Two-Sided (*p*-Value)	2.2×10^{-16} *	6.633×10^{-6} *	0.005486 *	0.3342	0.03193 *
One-Sided (*p*-Value)	2.2×10^{-16} *	3.317×10^{-6} *	0.002743 *	0.1679	0.01597 *

Note: * indicates that results are statistically significant based on *p*-value of 0.05.

Table 6. Kolmogorov-Smirnov predictive accuracy test (51 features).

	kNN vs. XGBoost	**Decision Tree vs. XGBoost**	**adaboostM1 vs. XGBoost**	**Random Forest vs. XGBoost**	**GBDT vs. XGBoost**
Two-Sided (*p*-Value)	2.2×10^{-16} *	5.228×10^{-5} *	0.00135 *	0.03609 *	0.01286 *
One-Sided (*p*-Value)	2.2×10^{-16} *	2.614×10^{-5} *	0.0006752 *	0.01805 *	0.006429 *

Note: * indicates that results are statistically significant based on *p*-value of 0.05.

Table 7. Kolmogorov-Smirnov predictive accuracy test (15 features).

	kNN vs. XGBoost	Decision Tree vs. XGBoost	adaboostM1 vs. XGBoost	Random Forest vs. XGBoost	GBDT vs. XGBoost
Two-Sided (*p*-Value)	2.2×10^{-16} *	0.0004882 *	0.01121 *	0.4522	0.0648
One-Sided (*p*-Value)	2.2×10^{-16} *	0.0002441 *	0.005604 *	0.2289	0.0324 *

Note: * indicates that results are statistically significant based on *p*-value of 0.05.

5. Discussion

Mammography is the primary breast imaging modality for early detection and diagnosis of breast cancer. However, achieving accurate diagnoses through mammography is often challenging for radiologists due to the difficulty in distinguishing the features of malignant lesions in dense breasts [34–36]. Consequently, a large amount of research is undertaken to develop computer-based applications, including various classification models [37–42].

Machine learning, especially on a large-scale for classifying breast cancer, remains an endeavor that is statistical in nature. Nevertheless, data obtained in this way can be associated with biomedical evidence. In this study, we employed one calcification dataset from one hospital. Our aim was to aid oncologists and medical image processing engineers in distinguishing benign from malignant breast cancers with high efficiency.

To date, various available machine learning methods have been used for identifying breast cancer. Jacob et al. [43] carried out a series of studies on various algorithms in the Wisconsin Breast Cancer diagnosis dataset. In kNN, an object is classified by a majority vote of its neighbors; namely, the object is assigned to the class most common among its k-nearest neighbors. AdaboostM1 is an ensemble algorithm that creates a highly accurate classifier by merging many relatively weak and inaccurate classifiers [44]. GBDT is an iterative decision tree algorithm composed of multiple decision trees. The results of all the trees are accumulated to provide the final result. GBDT is generally used for regression prediction. In this paper, we use it for classification after adjustment. XGBoost (extreme gradient boosting), first developed by Tianqi Chen and Gusetrin, is an open-source project. It is designed to implement an efficient, fast, scalable machine learning system (Gradient Tree Boosting) applicable to a wide variety of machine learning problems [9]. Here, we compared six popular techniques: kNN, decision tree, adaboostM1, RDF, GBDT, and XGBoost. We focused on the performance of XGBoost for the classification of breast cancer with microcalcifications. XGBoost had 90.24% accuracy in distinguishing benign from malignant lesions, achieving the best accuracy of all the other algorithms.

In addition, XGBoost has higher accuracy and lower false negatives compared with deep learning, although it lacks flexibility due to the requirement of manual feature extraction [45]. Recently, many studies identified relevant biomarkers or histopathological images for predicting diagnosis and outcome by XGBoost [46–52]. More importantly, XGBoost is effective in imaging for aiding the diagnosis of breast cancer. It has been reported that enhanced CT combined with XGBoost improves the performance of predicting the efficacy of anti-HER2 therapy for patients with liver metastases from breast tumor [53]. The integration of Ensemble Learning methods within mpMRI radiomic analysis helps in the diagnosis of breast cancer [54]. Radiomics and machine learning based on PET/CT images are used to predict HER2 status in breast cancer lesions [55]. Similarly, Vu et al. [56] found that the XGBoost model combined with clinical, mammographic, ultrasonographic, and histopathologic findings, assisted prognosis prediction in patients with breast cancer, reaching an accuracy of 0.84 and an AUC of 0.93. In this study, XGBoost was used to automatically discriminate between microcalcifications in mammograms, the main warning signs and even the only signs of breast cancer, yet with low sensitivity. Feature engineering is one of the important characteristics used in image classification. The majority of the traditional CAD systems rely on accurate features calculation for the microcalcification after

feature engineering [29]. In this study, 51 features were extracted according to the BI-RADS and were defined by the radiologists' requirements, which were clinically meaningful. In addition, the use of feature selection made a certain contribution to this study. The features ranked among the top 15 after feature selection are better interpreted by the clinicians, and were found to be promising and should be given more attention in clinical practice.

However, the current study suffered from the following limitations. First, this was a retrospective single-center study, and the sample was not conclusive. The testing and training dataset should be expanded and collected from different medical centers to achieve higher statistical power. In addition, the features extracted in our study may not be enough to fully characterize microcalcifications; thus, we will extract more meaningful ones in the future. By selecting and optimizing various features, it helps in the improvement of the performance of XGBoost in the classification stage. In future work, we will make a great effort to find a better representation of XGBoost and help in obtaining more describable information in breast cancer diagnosis. Moreover, this will further facilitate the systematic investigation of breast cancer for early detection, diagnosis, and clinical management [57].

6. Conclusions and Future Work

In this paper, we proposed an effective and efficient approach to the classification of breast cancer microcalcifications. This study finds a way to select the best discriminative features as a collection to improve the accuracy. It provides the best composition of features for efficient and accurate classification of breast cancers. With the set of specially selected features, we employed extreme gradient boosting to classify microcalcifications and achieved the highest accuracy of 90.24% on the dataset from our cancer center. This result demonstrates that it is essential for the classification of microcalcification to use the feature engineering method for the selection of the best composition of features, and the KSPA test results are statistically significant. Moreover, we showed that imaging segmentation makes a certain contribution to our research. In the future, we will make an effort to find a better representation of XGBoost, combined with feature engineering and selection, to obtain more describable information in breast cancer diagnosis.

Author Contributions: Writing—original draft, H.L.; Data curation, J.L., L.L. and X.J.; Validation, X.Z.; Writing—review & editing, H.W. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Meizhou Major Scientific and Technological Innovation Platforms and Projects of Guangdong Provincial Science&Technology Plan Projects under Grant No. 2019A0102005. This work was also funded by the Science and Technology Planning Project of Guangdong Province, China (No. 2021B1515310002).

Informed Consent Statement: We reviewed mammograms from 5476 female patients histopathologically diagnosed with benign or malignant lesions at the Sun Yat-sen University Cancer Center (SYSUCC, Guangzhou, China) between May 2011 and January 2017. The sample consisted of 2813 benign and 2663 malignant lesions. All patients underwent molybdenum-targeted mammography. All experimental protocols were approved by the Ethics Committee of SYSUCC and were conducted in accordance with Good Clinical Practice guidelines. Informed consent was obtained from each patient for their consent to have their information used in research without affecting their treatment option or violating their privacy.

Acknowledgments: We would like to thank all colleagues of Guangdong Key Laboratory of Big Data Analysis and Processing, Department of Computer Science, Sun Yat-sen University, and Sun Yat-sen University Cancer Center, who have worked in obtaining the tissue specimens and provided them for the current study.

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

References

1. Ferlay, J.; Soerjomataram, I.; Ervik, M.; Dikshit, R.; Eser, S.; Mathers, C.; Rebelo, M.; Parkin, D.M.; Forman, D.; Bray, F. *GLOBOCAN 2012 v1. 0, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11. 2013*; International Agency for Research on Cancer: Lyon, France, 2014.
2. Chen, W.; Zheng, R.; Baade, P.D.; Zhang, S.; Zeng, H.; Bray, F.; Jemal, A.; Yu, X.Q.; He, J. Cancer statistics in China, 2015. *CA A Cancer J. Clin.* **2016**, *66*, 115. [CrossRef] [PubMed]
3. Specht, J.M.; Mankoff, D.A. Advances in molecular imaging for breast cancer detection and characterization. *Breast Cancer Res.* **2012**, *14*, 206. [CrossRef] [PubMed]
4. Radiology ACo. *Breast Imaging Reporting and Data System Atlas (BI-RADS®Atlas)*; American College of Radiology: Reston, VA, USA, 2003.
5. Fletcher, S.W.; Elmore, J.G. Mammographic screening for breast cancer. *N. Engl. J. Med.* **2003**, *348*, 1672–1680. [CrossRef] [PubMed]
6. Cady, B.; Chung, M. Mammographic screening: No longer controversial: LWW. *Am. J. Clin. Oncol.* **2005**, *28*, 1–4. [CrossRef]
7. Lehman, C.D.; Lee, A.Y.; Lee, C.I. Imaging management of palpable breast abnormalities. *Am. J. Roentgenol.* **2014**, *203*, 1142–1153. [CrossRef] [PubMed]
8. Cheng, H.; Cai, X.; Chen, X.; Hu, L.; Lou, X. Computer-aided detection and classification of microcalcifications in mammograms: A survey. *Pattern Recognit.* **2003**, *36*, 2967–2991. [CrossRef]
9. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016.
10. Sajeev, S.; Bajger, M.; Lee, G. Superpixel texture analysis for classification of breast masses in dense background. *IET Comput. Vis.* **2018**, *12*, 779–786. [CrossRef]
11. Saleck, M.; ElMoutaouakkil, A.; Mouçouf, M. Tumor detection in mammography images using fuzzy C-means and GLCM texture features. In Proceedings of the 2017 14th International Conference on Computer Graphics, Imaging and Visualization, Marrakesh, Morocco, 23–25 May 2017; pp. 122–125.
12. Mohanty, A.K.; Senapati, M.R.; Lenka, S.K. An improved data mining technique for classification and detection of breast cancer from mammograms. *Neural Comput. Appl.* **2013**, *22*, 303–310. [CrossRef]
13. Zebari, D.A.; Ibrahim, D.A.; Zeebaree, D.Q.; Haron, H.; Salih, M.S.; Damaševičius, R.; Mohammed, M.A. Systematic Review of Computing Approaches for Breast Cancer Detection Based Computer Aided Diagnosis Using Mammogram Images. *Appl. Artif. Intell.* **2021**, *35*, 2157–2203. [CrossRef]
14. Dartois, L.; Gauthier, É.; Heitzmann, J.; Baglietto, L.; Michiels, S.; Mesrine, S.; Boutron-Ruault, M.; Delaloge, S.; Ragusa, S.; Clavel-Chapelon, F.; et al. A comparison between different prediction models for invasive breast cancer occurrence in the French E3N cohort. *Breast Cancer Res. Treat.* **2015**, *150*, 415–426. [CrossRef] [PubMed]
15. Cai, H.; Peng, Y.; Ou, C.; Chen, M.; Li, L. Diagnosis of breast masses from dynamic contrast-enhanced and diffusion-weighted MR: A machine learning approach. *PLoS ONE* **2014**, *9*, e87387. [CrossRef] [PubMed]
16. Krishnan, M.M.R.; Banerjee, S.; Chakraborty, C.; Chakraborty, C.; Ray, A.K. Statistical analysis of mammographic features and its classification using support vector machine. *Expert Syst. Appl.* **2010**, *37*, 470–478. [CrossRef]
17. Holsbach, N.; Fogliatto, F.S.; Anzanello, M.J. A data mining method for breast cancer identification based on a selection of variables. *Cienc. Saude Colet.* **2014**, *19*, 1295–1304. [CrossRef] [PubMed]
18. Şahan, S.; Polat, K.; Kodaz, H.; Güneş, S. A new hybrid method based on fuzzy-artificial immune system and k-nn algorithm for breast cancer diagnosis. *Comput. Biol. Med.* **2007**, *37*, 415–423. [CrossRef] [PubMed]
19. Friedman, J.H. Stochastic gradient boosting. *Comput. Stat. Data Anal.* **2002**, *38*, 367–378. [CrossRef]
20. Friedman, J.H.; Popescu, B.E. Importance sampled learning ensembles. *J. Mach. Learn. Res.* **2003**, *94305*, 1–32.
21. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
22. Friedman, J.; Hastie, T.; Tibshirani, R. Additive logistic regression: A statistical view of boosting (with discussion and a rejoinder by the authors). *Ann. Stat.* **2000**, *28*, 337–407. [CrossRef]
23. Moura, D.C.; López, M.A.G. An evaluation of image descriptors combined with clinical data for breast cancer diagnosis. *Int. J. Comput. Assist. Radiol. Surg.* **2013**, *8*, 561–574. [CrossRef]
24. Pérez, N.P.; López, M.A.G.; Silva, A.; Ramos, I. Improving the Mann–Whitney statistical test for feature selection: An approach in breast cancer diagnosis on mammography. *Artif. Intell. Med.* **2015**, *63*, 19–31. [CrossRef]
25. Arevalo, J.; González, F.A.; Ramos-Pollán, R.; Oliveira, J.L.; Lopez, M.A.G. Representation learning for mammography mass lesion classification with convolutional neural networks. *Comput. Methods Programs Biomed.* **2016**, *127*, 248–257. [CrossRef] [PubMed]
26. Pérez, N.; Guevara, M.A.; Silva, A. Improving breast cancer classification with mammography, supported on an appropriate variable selection analysis. In *Medical Imaging 2013: Computer-Aided Diagnosis*; International Society for Optics and Photonics: Bellingham, WA, USA, 2013; p. 22.
27. Pérez, N.; Guevara, M.A.; Silva, A.; Ramos, I.; Loureiro, J. Improving the performance of machine learning classifiers for Breast Cancer diagnosis based on feature selection. In Proceedings of the 2014 Federated Conference on Computer Science and Information Systems, Warsaw, Poland, 7–10 September 2014.
28. Clausi, D.A. An analysis of co-occurrence texture statistics as a function of grey level quantization. *Can. J. Remote Sens.* **2002**, *28*, 45–62. [CrossRef]

29. Haralick, R.M.; Shanmugam, K. Textural features for image classification. *IEEE Trans. Syst. Man Cybern.* **1973**, *3*, 610–621. [CrossRef]
30. Soh, L.-K.; Tsatsoulis, C. Texture analysis of SAR sea ice imagery using gray level co-occurrence matrices. *IEEE Trans. Geosci. Remote Sens.* **1999**, *37*, 780–795. [CrossRef]
31. Wei, X. *Gray Level Run Length Matrix Toolbox, v1. 0. Software*; Aeronautical Technology Research Center: Beijing, China, 2007.
32. Chu, A.; Sehgal, C.M.; Greenleaf, J.F. Use of gray value distribution of run lengths for texture analysis. *Pattern Recognit. Lett.* **1990**, *11*, 415–419. [CrossRef]
33. Hassani, H.; Silva, E.S. A Kolmogorov-Smirnov Based Test for Comparing the Predictive Accuracy of Two Sets of Forecasts. *Econometrics* **2015**, *3*, 590–609. [CrossRef]
34. Li, L.; Wu, Z.; Salem, A.; Chen, Z.; Chen, L.; George, F.; Kallergi, M.; Berman, C. Computerized analysis of tissue density effect on missed cancer detection in digital mammography. *Comput. Med. Imaging Graph.* **2006**, *30*, 291–297. [CrossRef]
35. Brem, R.F.; Hoffmeister, J.W.; Rapelyea, J.A.; Zisman, G.; Mohtashemi, K.; Jindal, G.; DiSimio, M.P.; Rogers, S.K. Impact of Breast Density on Computer-Aided Detection for Breast Cancer. *Am. J. Roentgenol.* **2005**, *184*, 439–444. [CrossRef]
36. Malich, A.; Marx, C.; Facius, M.; Boehm, T.; Fleck, M.; Kaiser, W.A. Tumour detection rate of a new commercially available computer-aided detection system. *Eur. Radiol.* **2001**, *11*, 2454–2459. [CrossRef]
37. Barlow, W.E.; Chi, C.; Carney, P.A.; Taplin, S.H.; D'Orsi, C.; Cutter, G.; Hendrick, R.E.; Elmore, J.G. Accuracy of Screening Mammography Interpretation by Characteristics of Radiologists. *JNCI J. Natl. Cancer Inst.* **2004**, *96*, 1840–1850. [CrossRef]
38. Muttarak, M.; Pojchamarnwiputh, S.; Chaiwun, B. Breast carcinomas: Why are they missed? *Singap. Med. J.* **2006**, *47*, 851.
39. Yu, S.; Guan, L. A CAD system for the automatic detection of clustered microcalcifications in digitized mammogram films. *IEEE Trans. Med. Imaging* **2000**, *19*, 115–126. [PubMed]
40. Jiang, Y.; Metz, C.E.; Nishikawa, R.M.; Schmidt, R.A. Comparison of Independent Double Readings and Computer-Aided Diagnosis (CAD) for the Diagnosis of Breast Calcifications. *Acad. Radiol.* **2006**, *13*, 84–94. [CrossRef] [PubMed]
41. Sankar, D.; Thomas, T. Fast fractal coding method for the detection of microcalcification in mammograms. In Proceedings of the 2008 International Conference on Signal Processing, Communications and Networking, Chennai, India, 4–6 January 2008.
42. Jiang, J.; Yao, B.; Wason, A. A genetic algorithm design for microcalcification detection and classification in digital mammograms. *Comput. Med. Imaging Graph.* **2007**, *31*, 49–61. [CrossRef] [PubMed]
43. Shomona Gracia, J.; Geetha Ramani, R. Efficient classifier for classification of prognostic breast cancer data through data mining techniques. In Proceedings of the World Congress on Engineering and Computer Science, San Francisco, CA, USA, 24–26 October 2012; 1.
44. Yoav, F.; Schapire, R.E. *Experiments with a New Boosting Algorithm*; ICML: Baltimore, MA, USA, 1996; pp. 148–156.
45. Arefan, D.; Mohamed, A.A.; Berg, W.A.; Zuley, M.L.; Sumkin, J.H.; Wu, S. Deep learning modeling using normal mammograms for predicting breast cancer risk. *Med. Phys.* **2020**, *47*, 110–118. [CrossRef]
46. Ai, H. GSEA–SDBE: A gene selection method for breast cancer classification based on GSEA and analyzing differences in performance metrics. *PLoS ONE* **2022**, *17*, e0263171. [CrossRef]
47. Thalor, A.; Joon, H.K.; Singh, G.; Roy, S.; Gupta, D. Machine learning assisted analysis of breast cancer gene expression profiles reveals novel potential prognostic biomarkers for triple-negative breast cancer. *Comput. Struct. Biotechnol. J.* **2022**, *20*, 1618–1631. [CrossRef]
48. Li, Q.; Yang, H.; Wang, P.; Liu, X.; Lv, K.; Ye, M. XGBoost-based and tumor-immune characterized gene signature for the prediction of metastatic status in breast cancer. *J. Transl. Med.* **2022**, *20*, 177. [CrossRef]
49. Jang, J.Y.; Ko, E.Y.; Jung, J.S.; Kang, K.N.; Kim, Y.S.; Kim, C.W. Evaluation of the Value of Multiplex MicroRNA Analysis as a Breast Cancer Screening in Korean Women under 50 Years of Age with a High Proportion of Dense Breasts. *J. Cancer Prev.* **2021**, *26*, 258–265. [CrossRef]
50. Jang, B.-S.; Kim, I.A. Machine-learning algorithms predict breast cancer patient survival from UK Biobank whole-exome sequencing data. *Biomark. Med.* **2021**, *15*, 1529–1539. [CrossRef] [PubMed]
51. Roy, S.; Das, S.; Kar, D.; Schwenker, F.; Sarkar, R. Computer Aided Breast Cancer Detection Using Ensembling of Texture and Statistical Image Features. *Sensors* **2021**, *21*, 3628. [CrossRef] [PubMed]
52. Chai, H.; Zhou, X.; Zhang, Z.; Rao, J.; Zhao, H.; Yang, Y. Integrating multi-omics data through deep learning for accurate cancer prognosis prediction. *Comput. Biol. Med.* **2021**, *134*, 104481. [CrossRef] [PubMed]
53. He, M.; Hu, Y.; Wang, D.; Sun, M.; Li, H.; Yan, P.; Meng, Y.; Zhang, R.; Li, L.; Yu, D.; et al. Value of CT-Based Radiomics in Predicating the Efficacy of Anti-HER2 Therapy for Patients With Liver Metastases From Breast Cancer. *Front. Oncol.* **2022**, *12*, 852809. [CrossRef] [PubMed]
54. Vamvakas, A.; Tsivaka, D.; Logothetis, A.; Vassiou, K.; Tsougos, I. Breast Cancer Classification on Multiparametric MRI – Increased Performance of Boosting Ensemble Methods. *Technol. Cancer Res. Treat.* **2022**, *21*, 15330338221087828. [CrossRef]
55. Chen, Y.; Wang, Z.; Yin, G.; Sui, C.; Liu, Z.; Li, X.; Chen, W. Prediction of HER2 expression in breast cancer by combining PET/CT radiomic analysis and machine learning. *Ann. Nucl. Med.* **2022**, *36*, 172–182. [CrossRef]
56. Vy, V.P.T.; Yao, M.M.-S.; Le, N.Q.K.; Chan, W.P. Machine Learning Algorithm for Distinguishing Ductal Carcinoma In Situ from Invasive Breast Cancer. *Cancers Basel.* **2022**, *14*, 2437. [CrossRef]
57. Wang, J.; Yang, X.; Cai, H.; Tan, W.; Jin, C.; Li, L. Discrimination of Breast Cancer with Microcalcifications on Mammography by Deep Learning. *Sci. Rep.* **2016**, *6*, 27327. [CrossRef]