Automated Quantitative Assessment of Amorphous Calcifications: Towards Improved Malignancy Risk Stratification

Kalyani Marathe[1], Chrysostomos Marasinou[2], Beibin Li[1], Noor Nakhaei[2], Bo Li[3], Joann G Elmore[4], Linda Shapiro[1], William Hsu[2]

[1] Department of Computer Science and Engineering, University of Washington, Seattle, WA

[2] Medical & Imaging Informatics, Department of Radiological Sciences, David Geffen School of Medicine at UCLA, Los Angeles, CA

[3] Department of Radiological Sciences, David Geffen School of Medicine at UCLA, Los Angeles, CA

[4] Department of Medicine, David Geffen School of Medicine at UCLA, Los Angeles, CA

Corresponding Author:

William Hsu, PhD

E-mail: whsu@mednet.ucla.edu

Phone: 310-794-3536

Fax: 310-794-3546

Abstract

Background. Amorphous calcifications noted on mammograms (i.e., small and indistinct calcifications that are difficult to characterize) are associated with high diagnostic uncertainty, often leading to biopsies. Yet, only 20% of biopsied amorphous calcifications are cancer. We present a quantitative approach for distinguishing between benign and actionable (high-risk and malignant) amorphous calcifications using a combination of local textures, global spatial relationships, and interpretable handcrafted expert features.

Method. Our approach was trained and validated on a set of 168 2D full-field digital mammography exams (248 images) from 168 patients. Within these 248 images, we identified 276 image regions with segmented amorphous calcifications and a biopsy-confirmed diagnosis. A set of local (radiomic and region measurements) and global features (distribution and expert-defined) were extracted from each image. Local features were grouped using an unsupervised k-means clustering algorithm. All global features were concatenated with clustered local features and used to train a LightGBM classifier to distinguish benign from actionable cases.

Results. On the held-out test set of 60 images, our approach achieved a sensitivity of 100%, specificity of 35%, and a positive predictive value of 38% when the decision threshold was set to 0.4. Given that all of the images in our test set resulted in a recommendation of a biopsy, the use of our algorithm would have identified 15 images (25%) that were benign, potentially reducing the number of breast biopsies.

Conclusions. Quantitative analysis of full-field digital mammograms can extract subtle shape, texture, and distribution features that may help to distinguish between benign and actionable amorphous calcifications.

Keywords: Radiomics, machine learning, mammography, microcalcifications

Introduction

Calcifications are a common mammographic finding. Radiologists use Breast Imaging, Reporting & Data Systems (BI-RADS) to report standardized qualitative descriptors of shape (e.g., popcorn-like, rod-like, round) and distribution (e.g., diffuse, segmental) to determine which calcifications are suspicious for malignancy. This risk stratification guides management decisions such as whom to recommend short-term imaging follow-up or biopsy. Calcifications that are too small and indistinct to assign a distinct shape are considered amorphous[1]. Prior studies have reported that the malignancy rate of biopsied amorphous calcifications is 20%[2].

While millions of women are encouraged to undergo mammography screening each year, radiologists continue to be challenged in evaluating and deciding which calcifications to biopsy. Unnecessary callbacks and biopsies lead to increased medical costs, patient anxiety, and potential morbidity. Studies have suggested that multiple descriptors for amorphous calcifications can lead to a higher positive predictive value (PPV) of malignancy[3]. However, the dependence on qualitative descriptors given by the radiologists is a limitation, given that many calcifications do not fit clearly into one morphologic or distribution category and manifest as a combination of descriptors[4]. Moreover, there is known inter and intraobserver variability in analyzing these calcifications[5].

Machine learning (ML) algorithms can potentially help overcome these limitations. Prior work on processing 2D full-field digital mammography (FFDM) has shown that calcifications can be detected and segmented with high sensitivity and accuracy[6–8]. A number of quantitative (radiomic) features can be calculated from these segmented calcifications to characterize the morphology, distribution, and texture of the calcifications and their surrounding regions. Radiomic features are defined using explicit formulas and are computed consistently across different images. Moreover, these features capture the subtle patterns that are difficult to assess visually by radiologists. Previous work has established that subtle differences in texture (e.g., differences in breast density) correlate with increased cancer risk[9]. Given this existing work, we present an AI/ML approach for utilizing radiomic features to predict their malignancy risk and inform whether further diagnostic workup is warranted. We use a combination

of radiomic and graph-theoretic features to distinguish between amorphous calcifications that are either benign (e.g., usually requiring only imaging follow-up) or actionable (e.g., requiring consideration of surgical intervention given the likelihood of having a high-risk/malignant lesion). We hypothesize that using more quantitative and precise descriptors of amorphous calcifications and the surrounding tissue can improve the PPV of identifying actionable (high-risk/malignant) lesions. Our contributions include 1) the characterization of amorphous calcifications using local textures, global spatial relationships, and other interpretable features; and 2) the use of unsupervised clustering to obtain a consistent set of local features independent of the number of calcifications across all images. We demonstrate how quantitative features generated from amorphous calcifications and their surrounding regions on 2D FFDM can help distinguish between women with benign findings versus those who require further workup (e.g., high-risk or malignant findings).

Materials and Methods

Data Collection

Following an Institutional Review Board-approved protocol, we collected a retrospective dataset with 1462 2D FFDM diagnostic exams performed at UCLA between 2017-2019. All cases had identified calcifications noted on the mammogram and underwent core breast biopsy. The presence of calcifications was identified using our breast screening registry (MagView, Fulton, MD). From this list, we selected the 359 exams with “amorphous calcification” findings mentioned in the radiology report and retrieved their corresponding images (n=2137) from our picture archive and communications system (PACS). A diagnostic exam may consist of images corresponding to different views (e.g., craniocaudal, mediolateral, exaggerated views). For this analysis, magnification views and images acquired using digital breast tomosynthesis were excluded, reducing the dataset by 71 exams. 87 exams with multiple pathology results (e.g., a case with malignant and high-risk lesions in the same breast) or non-amorphous calcifications upon re-review were omitted. After applying the exclusion criteria, a dataset with 178 diagnostic exams was generated consisting of 261 images with amorphous calcifications. Using the final radiology report, a trained researcher (CM), under the supervision of a fellowship-trained radiologist (BL), specified regions of interest (ROIs) on the FFDM pertaining to where the grouped amorphous calcifications were identified and ultimately biopsied. In total, 290 ROIs were annotated and matched with core-needle biopsy results. The breakdown of regions is as follows: 207 benign, 42 high-risk, 41 malignant. All images were acquired using a Hologic Selenia device at 0.07 mm per pixel resolution and 12-bit grayscale, with ~8.5 million pixels in each image.

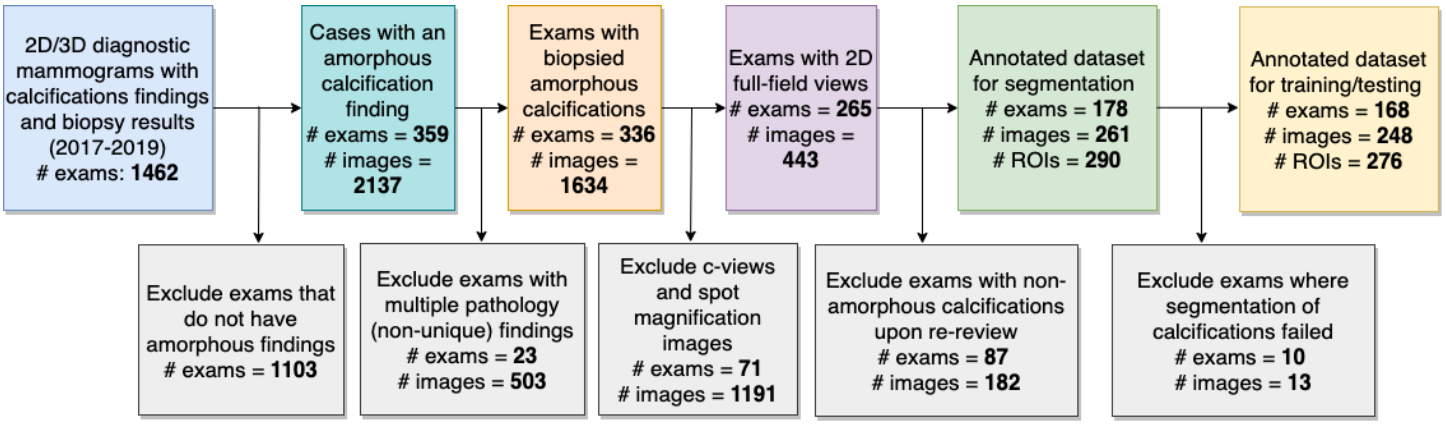

Figure 1: Cohort selection

Data Preparation

The input to our classification algorithm is a suspicious ROI showing amorphous microcalcifications that have been segmented. We executed the Hessian Difference of Gaussians Regression (HDoGReg) method to segment individual microcalcifications, a technique developed by Marasinou et al[10], on the entire 2D FFDM. The method consists of two stages: (1) bright candidate objects were delineated using difference-of-Gaussians blob detection with Hessian analysis for shape extraction, and (2) a convolutional regression model was applied to choose the candidate objects corresponding to microcalcifications. The resulting segmentation mask overlaps with the identified ROIs is used for classification analysis. 14 ROIs (13 images, 10 exams) did not overlap with any segmented calcifications and were omitted from our analysis. In total, 276 ROIs (248 images, 168 exams) were utilized for classification analysis. We split exams into two parts: 75% were assigned for training and 25% for testing. **Figure 1** summarizes the process for selecting and excluding exams. A breakdown of the dataset used to train and test our classifier is presented in **Table 1**.

Labels – (pathology outcome from breast biopsy)	**Train # ROIs (# images)**	**Test # ROIs (# images)**
Benign	154 (131 images)	46 (43 images)
Actionable (High-risk/DCIS/invasive)	57 (57 images)	19 (17 images)
Total	211 (188 images)	65 (60 images)

Table 1: Breakdown in the training and testing set reported as # of ROIs (# of images)

Overall approach

An overview of the classification pipeline is shown in **Figure 2**. Three types of binary masks were generated using the inputted ROIs with segmented microcalcifications: foreground, background, and dilated foreground, as explained below. From each of these masks, local and global features were extracted. Local features quantify the shape and texture of individual microcalcifications and their immediate surrounding regions. These local features are then aggregated using a k-means clustering algorithm to create a fixed-dimensional feature vector per image to characterize their distributions.

Global (distribution and expert-defined) features, calculated from the foreground and dilated foreground masks, characterize the overall distribution of the microcalcifications within an ROI. All global features were concatenated with clustered local features and utilized to train a LightGBM classifier to distinguish benign from actionable cases. The model was tuned to achieve high sensitivity given the clinical importance of catching all potential high-risk and malignant lesions. Classification metrics of the final model were reported.

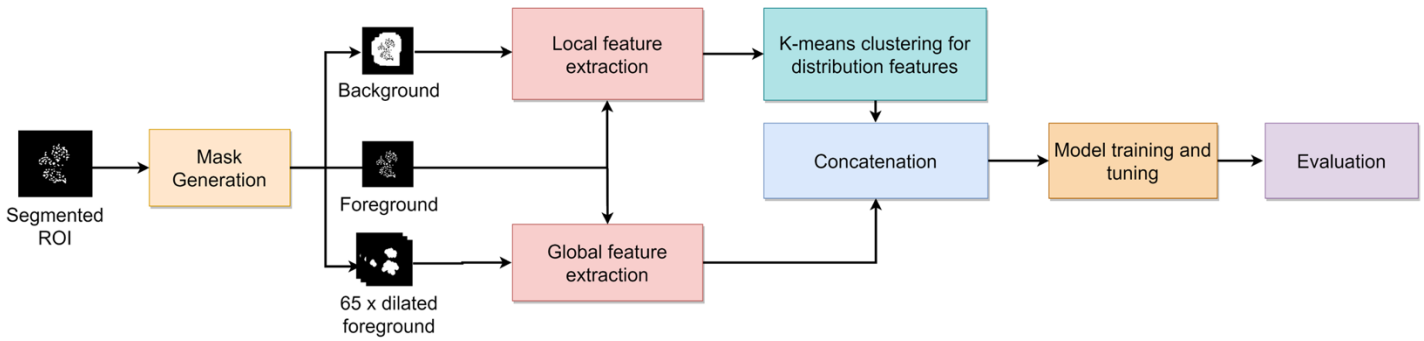

Figure 2: Classification pipeline

Mask Generation

We applied the microcalcification segmentation algorithm[10] to generate three different types of segmentation masks, resulting in 10 masks per ROI:

- **1 x foreground mask**, individual microcalcifications are identified in the foreground mask. An example is shown in **Figure 3(b).**
- **1 x background mask**, a 25-pixel band surrounding each microcalcification to capture the surrounding tissue. **Figure 3(c)** provides an example. The thickness of the layer was determined based on feedback from expert radiologists on the relevance of surrounding breast tissue in informing the diagnosis.
- **65x dilated foreground masks**, a morphological dilation of the foreground mask at eight different scales (e.g., 1x-65x). Dilated mask examples at two different scales are shown in **Figure 4**. These dilated masks are used to determine the spatial relationships among groups of calcifications when computing the topological features.

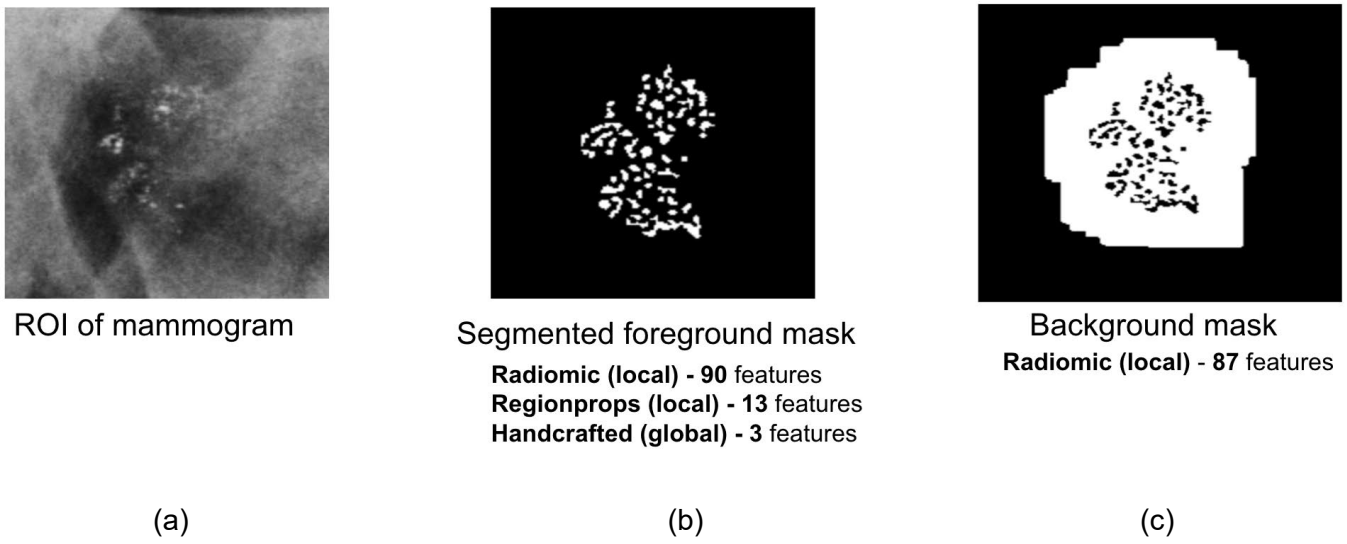

Figure 3: (a) An example ROI, (b) the foreground mask, and (c) the background mask.

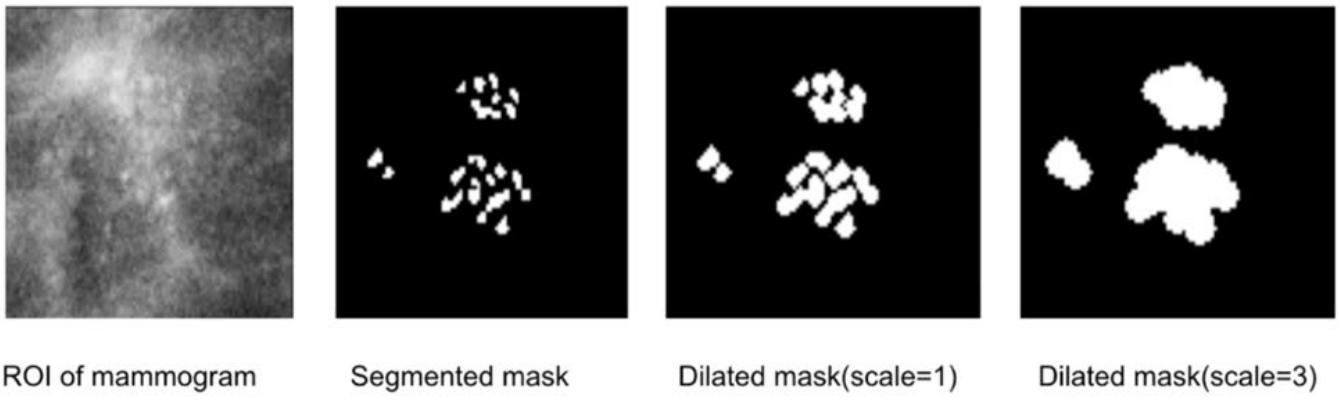

Figure 4: Visualizations of dilated foreground masks at two representative scales

Local Features

Using the generated foreground and background masks, radiomic features (e.g., intensity, shape, texture) and region properties (e.g., area, intensity) of each labeled region were then extracted to create three sets of features, which are enumerated in the **Supplementary Materials**. From the foreground mask, which consists of individually segmented microcalcifications, 90 radiomic features (using pyradiomics[11]) and 13 region measurements (using regionprops module of scikit-image[12]) were generated. From the background mask, which corresponds to the breast parenchyma immediately surrounding the calcification group, 87 radiomic features (using pyradiomics[11]) were computed.

Global Features

Region-level features were extracted using the foreground mask and the eight dilated foreground masks. The features characterize the distribution of microcalcifications and their topological structure. In total, 67 global features per ROI were extracted as described below.

Multiscale topological features: Following the work of Chen et al[13], we computed eight features describing the distribution of microcalcifications at 65 different scales, zero to 64 dilation scales, resulting in a total of 520 features. Then using feature importance, 122 features were selected. In the next stage, first, the connectivity graphs between individual calcifications were constructed. Each calcification in the foreground mask represents a node in the graph. For each of the dilated foreground masks, overlapping objects due to the dilation operation were considered connected, and a graph vertex was drawn between them, leading to the generation of 8 graphs per ROI. Then, for each graph, eight topological features were extracted: 1) number of subgraphs, 2) average vertex degree, 3) maximum vertex degree, 4) average vertex eccentricity, 5) diameter, 6) average clustering coefficient, 7) giant connected component ratio, and 8) the percentage of isolated points. The formulae for computing these features are given in the **Supplementary Materials**.

Handcrafted features: Based on input from a fellowship-trained breast radiologist (BL) and comparison with deep learning-based feature extractors, we considered three additional types of features:

1. Standard Deviation of Area: the microcalcifications that vary in size and shape tend to be considered highly suspicious for malignancy[14]. To quantify the variability, we calculate the standard deviation of the areas of individual microcalcifications within ROIs of each image.
2. Correlation Coefficient: the microcalcifications' patterns are crucial in determining whether they are suspicious or not. Since most malignancies are ductal, the linear distribution patterns of microcalcifications suggest that the patient needs further follow-up[15]. Therefore, we calculated the correlation coefficient of the x and y coordinates of the centroids of microcalcifications from ROIs of each image to quantify the extent of their linear distribution.
3. Pairwise distances: calcifications that are spread over a large volume or over the entire breast are more likely to be benign[14]. To quantify the spread, we computed the pairwise mean distance of the microcalcifications in ROIs.

Clustering and concatenation

The three local feature sets (foreground radiomics, foreground region properties, and background radiomics) are used as inputs into the classifier but are proportional to the number of objects (microcalcifications and background regions) in each image, which varies. To obtain a consistent set of features for each image, we apply an unsupervised approach for aggregating local features in each feature set to represent their distribution as a fixed-dimensional vector. An unsupervised K-means clustering was utilized to group individual microcalcifications or background regions with similar characteristics. All objects within an ROI were labeled using an integer representing their K-means cluster for each image. After counting the number of

microcalcifications and background regions in each cluster, a K-dimensional feature vector is then constructed where each element represents the percentage of objects belonging to a particular cluster. This process was carried out for each of the three local feature sets. The K-means clustering model was fitted on the training data. As an alternative to K-means clustering, we represented the distribution of local features by computing the mean and standard deviation of each feature value across all objects.

The three vectors representing local feature sets were then concatenated with the global features (distribution and handcrafted features) to form an image's final feature vector. An illustration of the aggregation method of one feature set is shown in **Figure 5**.

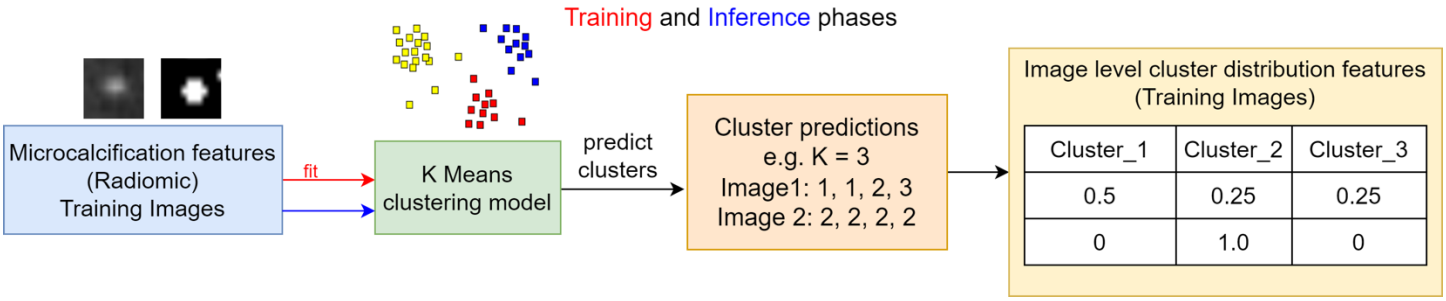

Figure 5: K means clustering-based feature aggregation pipeline. During training, we generated clusters using the features extracted from the objects of training ROIs. During testing, we utilized the clusters created during the training phase to predict the clusters of the objects from the testing ROIs. The process was repeated to generate three local feature sets followed by their concatenation with global features.

Model training and evaluation

The LightGBM classifier[15] was used to perform a grid search with five-fold cross-validation on the training data to identify the best hyperparameters for this classification task. To address the issue of class imbalance, SMOTE + Edited Nearest Neighbor resampling technique was used before training[16]. Along with the hyperparameters (regularization, number of trees, learning rate, number of leaves, max depth), the probability threshold, and the optimal number of clusters, K was also tuned on the training set to obtain the best possible sensitivity to ensure that we do not miss any cancerous cases. The optimal number of clusters, K, was determined using five-fold cross-validation to maximize the classification sensitivity, yielding K=15 and a decision threshold of 0.4. The model was retrained using the chosen hyperparameters on the entire training set, and the images in the test set were classified. Metrics such as accuracy, sensitivity, specificity, F1, PPV, and receiver operating characteristic (ROC) curve area under the curve (AUC) are reported.

We also compared our radiomics-based clustering approach against three alternatives: 1) transfer learning by fine-tuning a ResNet-50 model pre-trained using ImageNet with weighted cross-entropy loss; 2) a LightGBM classifier with features extracted from the fine-tuned ResNet-50 model, and 3) a radiomic feature-based approach with statistical features (i.e., mean and standard deviation) computed across all microcalcifications instead of clustering.

For fine-tuning the ResNet-50 pre-trained with ImageNet, we split the training exams into 80% training and 20 % validation. The model that gave the best F-1 on the validation data was used for evaluation. To address the problem of class imbalance, we used weighted cross-entropy loss as an objective function and data augmentations to avoid overfitting. However, during the process of fine-tuning, the ROC AUC of the training data showed an increasing trend while the ROC AUC of the validation data did not improve. This observation demonstrated that the model was overfitted to our training data and failed to perform well in a data-scarce scenario such as this.

To estimate an unbiased generalization performance of our algorithm, we performed nested cross-validation on the entire dataset. The outer loop of the nested cross-validation estimates the model performance, while the

inner loop is used for hyperparameter tuning using grid search. For the outer loop, in addition to the results of the single 75% training-25% testing split reported previously, we re-ran our entire analysis on the remaining three stratified splits. For the inner loop, the training data was further divided into five stratified folds (80% training, 20% testing), out of which one fold was used for validation, and the rest of the folds were used for training. The best parameters obtained from the grid search were then used to train the final model on the training data and then evaluated on the test data. This procedure was carried out on all four splits, and the averaged results were calculated, assessing whether the clustering approach consistently outperforms the alternative methods.

Results

Sensitivity and specificity

On the held-out test set of 60 images, we obtained 100% sensitivity when the decision threshold was 0.4 (**Table 2**). The PPV was 38% due to the high number of false positives. Specificity can be improved by increasing the decision threshold. The values of accuracy, f1, and specificity for alternative decision thresholds are shown in **Supplemental Materials**. Compared to using the mean and standard deviation of the amorphous calcification features, the clustering approach is superior, though both methods achieve a sensitivity of 100%.

	Clustering Approach (K = 15 and probability threshold = 0.4)	**Approach using Mean and Standard Deviation of amorphous calcification features** (Probability threshold = 0.4)
Accuracy	0.53	0.31
Sensitivity	**1.0**	**1.0**
Specificity	0.35	0.04
F1	0.51	0.08
PPV	0.38	0.29
ROC AUC	0.73	0.55

Table 2: Classification results - Clustering approach versus using mean and standard deviation

ROC analysis

Evaluating using the independent test set, the area under the ROC curve (shown in **Figure 6**) is 0.73 classifying an ROI as either benign or actionable using the clustering approach. The clustering approach outperformed the approach using local features' mean and standard deviation to create global features (ROC AUC = 0.55).

Recent advances in machine learning have yielded deep feature extractors capable of automatically learning informative features from the data rather than handcrafted features. To compare the performance of a model using deep features, we conducted three experiments in which we ran our classification pipeline using (a) our local and global features and (b) the 2048 features from the last layer of the fine-tuned ResNet-50 model. (c) the

4096 features from the last layer of the fine-tuned VGG-16 model. While the specificity of the ResNet-50-based approach was comparable to our clustering approach (0.35 for the clustering approach versus 0.37 for the fine-tuned ResNet), the sensitivity (1.0) and ROC AUC (0.73) achieved by our clustering approach were superior compared to the sensitivity (0.89) and ROC AUC (0.58) that was obtained using fine-tuned ResNet-50 features at the probability threshold of 0.4. We note that the ResNet model was overfitted during training and performed poorly during testing, even with data augmentation, early stopping, and other regularization methods like weight decay having been applied. The VGG-16 based features also followed a similar pattern and achieved a sensitivity of 0.95 and ROC AUC of 0.52, which is lower than our clustering approach. The specificity of 0.12 compared to 0.35 (achieved using our clustering approach).

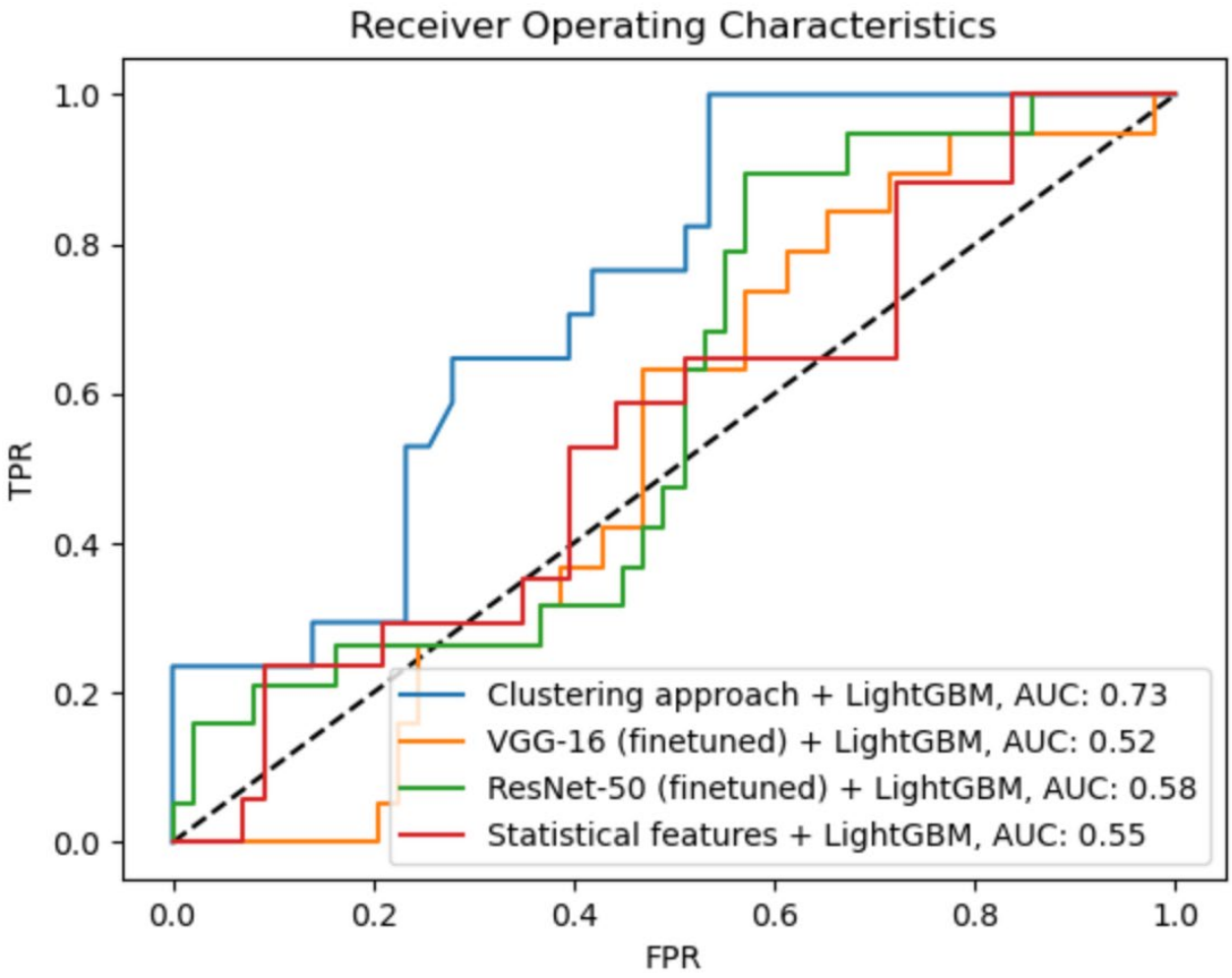

Figure 6: ROC curve of the classification using (a) K-means clustering-based aggregation of local textural features and global features with LightGBM classifier (b) Features extracted from fine-tuned VGG-16 using weighted cross-entropy loss+ LightGBM classifier (c) Features extracted from fine-tuned ResNet-50 using weighted cross-entropy loss + LightGBM classifier (d) Mean and standard deviation aggregation of local features + LightGBM classifier

Confusion matrix

The confusion matrices are shown in **Table 3**, comparing the clustering and mean and standard deviation approaches. For the clustering approach, all 17 images were correctly classified as actionable and 15 classified as benign for probability threshold = 0.4 with our clustering approach. We chose a threshold that emphasized higher sensitivity (e.g., not missing any potential cancers) at the cost of an increased number of false positives (e.g., obtaining biopsies on benign findings).

	Clustering approach		**Approach using Mean and standard deviation**	
	Benign (Predicted)	Actionable (Predicted)	Benign (Predicted)	Actionable (Predicted)
Benign	15	28	2	41
Actionable	0	17	0	17

Table 3: Confusion matrix for clustering approach and probability threshold of 0.4 and approach using mean and standard deviation.

Generalization performance of our approach

Reporting results from the full nested cross-validation, our clustering approach remained the best performing approach, a mean ROC AUC of 0.71±0.12 compared to: 1) a pre-trained VGG-16 (not fine-tuned) + LightGBM approach, resulting in an ROC AUC of 0.54 ± 0.1, 2) a finetuned VGG-16 model resulting in an ROC AUC of 0.48 ± 0.04, 3) a finetuned VGG-16 + LightGBM approach 0.52 ± 0.03 4) a pre-trained ResNet-50 (not fine-tuned) + LightGBM approach resulting in an ROC AUC of 0.47 ± 0.07, 5) a fine-tuned ResNet-50 model, resulting in an ROC AUC of 0.42±0.07, 6) a ResNet-50 + LightGBM approach, resulting in an ROC AUC of 0.48±0.07, and 7) a radiomic feature-based approach with statistical features, resulting in an ROC AUC of 0.53±0.06.

Discussion

Amorphous microcalcifications on mammography images are challenging for radiologists to assess and lead to a high number of biopsies of benign findings. Quantitative analysis of the morphology and distribution of amorphous microcalcifications has the potential to better distinguish between benign and actionable findings. In this analysis, we demonstrate that in a challenging subset of cases that were all referred for biopsy, the algorithm correctly identified 15/60 (25%) benign images, potentially saving these women from undergoing unnecessary breast biopsies. Moreover, the algorithm using unsupervised clustering achieved a 38% PPV compared to a PPV of 28% that radiologists achieved on these images. In our test set, perfect sensitivity was achieved in identifying all actionable findings, but at the continued cost of many false positives despite the improvement in PPV.

Several related studies on classifying different calcifications have been previously published, but none focused on developing and testing algorithms for challenging amorphous calcifications. Fanizzi et al[17] utilized SURF

(Speeded Up Robust Features) to detect a range of calcifications and extract wavelet decomposition features from the surrounding regions on screening digital mammograms. Trained on 130 ROIs (75 benign, 55 malignant) and tested on another 130 ROIs, they reported 0.92 ROC AUC, 88% accuracy, 87% sensitivity, and 88% specificity to classify microcalcifications that are associated with benign/malignant lesions. Karahaliou et al[18] utilized 85 full-field digitized screen-film images originating from the Digital Database for Screening Mammography (DDSM) and extracted 128 x 128 ROIs centered around detected microcalcification clusters. They reported a ROC AUC of 0.84, and given a threshold that optimizes sensitivity, they achieved 94.4% sensitivity but with a high false positive rate (20% specificity). We achieve a similar sensitivity but much higher specificity in our work on the subset of more challenging smaller amorphous microcalcifications. Finally, Stelzer et al[19] manually segmented various calcifications from magnification views of diagnostic FFDMs and extracted 249 features from 235 cases with stereotactic biopsy-proven diagnoses. They showed that 37–46% of biopsies could be avoided per reader at the cost of one false-negative. Focusing on the clinically challenging amorphous calcifications, our method could avoid biopsying 25% of the images with no false negatives.

To identify the most informative features for this classification task, we examined the most represented features in the trees used to build the classifier. These features include 1) region property distributions of foreground regions (i.e., area, perimeter, axes lengths, the solidity of the microcalcification regions); 2) graph-based feature clusters (i.e., calcification distribution); 3) pairwise mean distance (i.e., the spread of the microcalcifications); and 4) standard deviation of calcification size (i.e., variation in the calcification size).

Several limitations of our work are noted. First, the input dataset consists of 261 annotated images, which is limited but similar to previously reported studies[17–19]. The number of images is also constrained due to our focus on amorphous calcifications. This subset of calcifications is associated with the greatest diagnostic uncertainty and a high false positive rate of breast biopsies. Second, our analysis only included cases where the segmentation algorithm generated a result: images in which the segmentation algorithm detected no objects or only outputted a single object were excluded. This assumption may introduce a source of bias. Third, given the small and hazy nature of amorphous calcifications, the segmentation algorithm could identify false positive objects that may impact the accuracy of the classifier.

Further analysis involving a larger number of images and data from external institutions and mammography devices is needed to investigate the robustness and generalizability of the pipeline. Fourth, the pre-trained VGG and fine-tuned ResNet models' performance was suboptimal due to overfitting. While a more diverse training set could improve their performance, our clustering approach achieves consistently better performance despite the limited sample size.

In summary, this work provides initial evidence that a quantitative approach to characterizing amorphous microcalcifications noted on mammography examinations, generating information about the shape and distribution of each calcification, can improve the ability to distinguish between benign and actionable findings. Our algorithm identified as benign 25% of microcalcifications that were originally deemed suspicious by the radiologists (and leading to a breast biopsy), potentially decreasing the number of false positive biopsies.

Acknowledgments

The authors would like to acknowledge funding support from the National Cancer Institute R37 CA240403 to J.E. and W.H and U01 CA231782 to B.L., L.S., and J.E. and from the National Science Foundation grant number 1722516 to C.M. and W.H.

References

1 Bassett LW, Lee-Felker S. 26 - Breast Imaging Screening and Diagnosis. In: Bland KI, Copeland EM, Klimberg VS, Gradishar WJ, editors. *Breast Fifth Ed.* Elsevier; 2018. p. 337-361.e2.

2 Moy L. Should We Continue to Biopsy All Amorphous Calcifications? *Radiology* 2018;**288**:680–1. https://doi.org/10.1148/radiol.2018180767.

3 Oligane HC, Berg WA, Bandos AI, Chen SS, Sohrabi S, Anello M, *et al.* Grouped Amorphous Calcifications at Mammography: Frequently Atypical but Rarely Associated with Aggressive Malignancy. *Radiology* 2018;**288**:671–9. https://doi.org/10.1148/radiol.2018172406.

4 Berg WA, Arnoldus CL, Teferra E, Bhargavan M. Biopsy of Amorphous Breast Calcifications: Pathologic Outcome and Yield at Stereotactic Biopsy. *Radiology* 2001;**221**:495–503. https://doi.org/10.1148/radiol.2212010164.

5 Lee AY, Wisner DJ, Aminololama-Shakeri S, Arasu VA, Feig SA, Hargreaves J, *et al.* Inter-reader Variability in the Use of BI-RADS Descriptors for Suspicious Findings on Diagnostic Mammography: A Multi-institution Study of 10 Academic Radiologists. *Acad Radiol* 2017;**24**:60–6. https://doi.org/10.1016/j.acra.2016.09.010.

6 Ciecholewski M. Microcalcification Segmentation from Mammograms: A Morphological Approach. *J Digit Imaging* 2017;**30**:172–84. https://doi.org/10.1007/s10278-016-9923-8.

7 El-Naqa I, Yongyi Yang, Wernick MN, Galatsanos NP, Nishikawa R. Support Vector Machine Learning for Detection of Microcalcifications in Mammograms. Presented at the 2002 IEEE International Symposium on Biomedical Imaging, Washington, DC, USA.

8 Wang J, Yang Y. A context-sensitive deep learning approach for microcalcification detection in mammograms. *Pattern Recognit* 2018;**78**:12–22. https://doi.org/10.1016/j.patcog.2018.01.009.

9 Kontos D, Winham SJ, Oustimov A, Pantalone L, Hsieh M-K, Gastounioti A, *et al.* Radiomic Phenotypes of Mammographic Parenchymal Complexity: Toward Augmenting Breast Density in Breast Cancer Risk Assessment. *Radiology* 2019;**290**:41–9. https://doi.org/10.1148/radiol.2018180179.

10 Marasinou C, Li B, Paige J, Omigbodun A, Nakhaei N, Hoyt A, *et al.* Segmentation of Breast Microcalcifications: A Multi-Scale Approach. *ArXiv210200754 Eess* 2021.

11 van Griethuysen JJM, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, *et al.* Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer Res* 2017;**77**:e104–7. https://doi.org/10.1158/0008-5472.CAN-17-0339.

12 van der Walt S, Schönberger JL, Nunez-Iglesias J, Boulogne F, Warner JD, Yager N, *et al.* scikit-image: image processing in Python. *PeerJ* 2014;**2**:e453. https://doi.org/10.7717/peerj.453.

13 Chen Z, Strange H, Oliver A, Denton ER, Boggis C, Zwiggelaar R. Topological modeling and classification of mammographic microcalcification clusters. *IEEE Trans Biomed Eng* 2014;**62**:1203–14.

14 Nalawade YV. Evaluation of breast calcifications. *Indian J Radiol Imaging* 2009;**19**:282–6. https://doi.org/10.4103/0971-3026.57208.

15 Ke, Guolin, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. "Lightgbm: A highly efficient gradient boosting decision tree." Advances in neural information processing systems 30 (2017): 3146-3154.

16 Douzas G, Bacao F. Geometric SMOTE a geometrically enhanced drop-in replacement for SMOTE. *Inf Sci* 2019;**501**:118–35. https://doi.org/10.1016/j.ins.2019.06.007.

17 Fanizzi A, Basile TMA, Losurdo L, Bellotti R, Bottigli U, Dentamaro R, *et al.* A machine learning approach on multiscale texture analysis for breast microcalcification diagnosis. *BMC Bioinformatics* 2020;**21**:91. https://doi.org/10.1186/s12859-020-3358-4.

18 Karahaliou A, Skiadopoulos S, Boniatis I, Sakellaropoulos P, Likaki E, Panayiotakis G, *et al.* Texture analysis of tissue surrounding microcalcifications on mammograms for breast cancer diagnosis. *Br J Radiol* 2007;**80**:648–56. https://doi.org/10.1259/bjr/30415751.

19 Stelzer PD, Steding O, Raudner MW, Euller G, Clauser P, Baltzer PAT. Combined texture analysis and machine learning in suspicious calcifications detected by mammography: Potential to avoid unnecessary stereotactical biopsies. *Eur J Radiol* 2020;**132**:109309. https://doi.org/10.1016/j.ejrad.2020.109309.