cancers

MDPI

Article

Temporal Machine Learning Analysis of Prior Mammograms for Breast Cancer Risk Prediction

Hui Li [1], Kayla Robinson [1], Li Lan [1], Natalie Baughan [1], Chun-Wai Chan [1], Matthew Embury [2], Gary J. Whitman [3], Randa El-Zein [4], Isabelle Bedrosian [2,*] and Maryellen L. Giger [1,*]

[1] Department of Radiology, The University of Chicago, Chicago, IL 60637, USA; huili@uchicago.edu (H.L.)
[2] Department of Breast Surgical Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA
[3] Department of Breast Imaging, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA
[4] Department of Radiology, Houston Methodist Research Institute, Houston, TX 77030, USA
* Correspondence: ibedrosian@mdanderson.org (I.B.); m-giger@uchicago.edu (M.L.G.)

Simple Summary: Machine learning approaches, using both radiomic and deep-learning-based features, were performed for an analysis of the breast parenchyma to identify women at risk of future breast cancer. Results from this study demonstrate that the antecedent mammographic images can potentially discriminate between women with a future-biopsy-proven cancer versus those with a future-biopsy-proven benign lesion.

Abstract: The identification of women at risk for sporadic breast cancer remains a clinical challenge. We hypothesize that the temporal analysis of annual screening mammograms, using a long short-term memory (LSTM) network, could accurately identify women at risk of future breast cancer. Women with an imaging abnormality, which had been biopsy-confirmed to be cancer or benign, who also had antecedent imaging available were included in this case–control study. Sequences of antecedent mammograms were retrospectively collected under HIPAA-approved guidelines. Radiomic and deep-learning-based features were extracted on regions of interest placed posterior to the nipple in antecedent images. These features were input to LSTM recurrent networks to classify whether the future lesion would be malignant or benign. Classification performance was assessed using all available antecedent time-points and using a single antecedent time-point in the task of lesion classification. Classifiers incorporating multiple time-points with LSTM, based either on deep-learning-extracted features or on radiomic features, tended to perform statistically better than chance, whereas those using only a single time-point failed to show improved performance compared to chance, as judged by area under the receiver operating characteristic curves (AUC: 0.63 $\pm$ 0.05, 0.65 $\pm$ 0.05, 0.52 $\pm$ 0.06 and 0.54 $\pm$ 0.06, respectively). Lastly, similar classification performance was observed when using features extracted from the affected versus the contralateral breast in predicting future unilateral malignancy (AUC: 0.63 $\pm$ 0.05 vs. 0.59 $\pm$ 0.06 for deep-learning-extracted features; 0.65 $\pm$ 0.05 vs. 0.62 $\pm$ 0.06 for radiomic features). The results of this study suggest that the incorporation of temporal information into radiomic analyses may improve the overall classification performance through LSTM, as demonstrated by the improved discrimination of future lesions as malignant or benign. Further, our data suggest that a potential field effect, changes in the breast extending beyond the lesion itself, is present in both the affected and contralateral breasts in antecedent imaging, and, thus, the evaluation of either breast might inform on the future risk of breast cancer.

Keywords: breast cancer risk; radiomics; long short-term memory networks; artificial intelligence; field effect

Citation: Li, H.; Robinson, K.; Lan, L.; Baughan, N.; Chan, C.-W.; Embury, M.; Whitman, G.J.; El-Zein, R.; Bedrosian, I.; Giger, M.L. Temporal Machine Learning Analysis of Prior Mammograms for Breast Cancer Risk Prediction. *Cancers* **2023**, *15*, 2141. https://doi.org/10.3390/cancers15072141

Academic Editors: Tommaso Susini and Laura Papi

Received: 17 February 2023
Revised: 24 March 2023
Accepted: 29 March 2023
Published: 4 April 2023

1. Introduction

While agencies such as the American College of Radiology, American College of Physicians, and American Cancer Society have different recommendations for breast screening frequency guidelines, they all suggest mammographic screening with some frequency over some portion of a woman's lifetime [1–3]. Women who follow these guidelines produce, over the years, temporal sequences of mammographic images. When interpreting screening exams, radiologists often compare current mammograms with prior mammograms to qualitatively assess the interval change in breast tissue. Such a practice is conducted because the interval change may indicate the development of a new cancer [4].

It has been demonstrated that comparing current and prior mammograms improves specificity in breast cancer screening. A study that compared performance on over one million images found that the use of comparison mammograms at screening resulted in lower recall rates (6.9% with comparison mammograms vs. 14.9% without comparison mammograms) and higher specificity (93.5% with comparison mammograms vs. 85.7% without comparison mammograms) [4]. This suggests that in ambiguous cases, where it is not obvious whether an abnormality poses a threat, the changes in mammograms over time provide the radiologist with discriminatory information that helps inform the decision of whether or not to send a patient for follow-up. For example, if a suspicious region is judged to be visible and unchanged from prior mammograms, then the risk of malignancy may be lower as evaluated by the radiologist. The utility of prior images in radiologist review suggests the incorporation of prior images may also be informative in artificial-intelligence-based cancer prediction systems aimed at assisting the radiologist in detecting cancer risk.

A number of studies have shown the utility of incorporating prior imaging exams in clinical classification tasks. A study by Santeramo et al. [5] implemented a time-modulated long short-term memory (LSTM) network to detect abnormalities in a database of 745,480 chest X-rays, with the intent to classify abnormalities as either cardiomegaly, consolidation, pleural effusion, or hiatus hernia. The study compared the performance of a convolutional neural network (CNN Inception v3) trained on single images as a baseline to an LSTM network using the single images plus prior longitudinal observations. Using the F-measure as a figure of merit, the study observed, on average over the four abnormality types, that the LSTM resulted in a 7% increase in F-measure and a 9% increase in PPV over the baseline, single-image CNN. A study by Shao et al. [6] investigated the use of temporal radiomics to interrogate normal appearing white matter (NAWM) in order to predict the development of white matter hyperintensities (WMH) which are associated with cognitive decline among elderly patients. This study constructed radiomic signatures on regions of interest among a cross sectional cohort of cases with noted progression of WMH and aged-matched controls without progression to WMH, each of which had undergone two or more MRI exams on the same scanner with a time period of at least one year between scans. The study reported an area under the curve (AUC) of 0.954 (95% confidence interval: 0.876–0.989) for distinguishing between areas of NAWM that developed into WMH from those that did not develop into WMH. In addition, in predicting the malignancy of breast lesions on dynamic contrast-enhanced magnetic resonance images (MRI), LSTM has been used to incorporate the multiple acquisition time-points within the dynamic imaging protocol [7]. Specifically, Antropova et al. demonstrated higher classification performance on lesion characterization with MRI using LSTM than using a fine-tuned feed-forward network at a single time-point [7]. These studies provide evidence that a computer analysis of temporal images may improve the accuracy of predicting future disease.

Given the relevance of serial imaging in the diagnostic interpretation of mammographic findings and the emerging findings on the importance of incorporating temporal data for the classification of a future disease state, we sought to test the hypothesis that a computer analysis of multiple sequential antecedent mammograms could predict the future risk of benign versus malignant breast lesions. Our study investigates both conventional

human-engineered radiomic features and deep learning methods for the task of classifying future lesions.

In order to incorporate information collected over a time series of full field digital mammograms (FFDMs), we chose to use an LSTM network in this study, as it is capable of learning long-term dependencies for data organized as a series [8]. As a recurrent neural network (RNN), LSTM networks are able to retain information about previous time-points in a series and use this information to inform decisions on the present time-points of that same series [9,10]. LSTM networks can take in feature vectors from various sources, and so this study explored the performance of an LSTM trained on features extracted from a CNN and the performance of an LSTM trained on conventional human-engineered features extracted from the same images. Additionally, we measured the performance obtained by extracting features from a single time-point and merging features using a support vector machine (SVM) classifier. In this way, an assessment was performed between deep features and conventional human-engineered features as well as between time series data and single-time-point data for classification.

2. Materials and Methods

2.1. Image Acquisition and Database Description

Mammograms were retrospectively collected from MD Anderson Cancer Center and the University of Chicago Medical Center of women who had undergone screening exams for two or more years prior to the detection of a mammographic abnormality. Subjects identified at MD Anderson were part of a cohort of women recruited prospectively evaluating blood and tissue biomarkers of breast cancer risk; the subset of subjects with prior mammograms was included in this analysis. Subjects at the University of Chicago were identified retrospectively from an imaging database of women undergoing both screening and diagnostic mammograms. Images were acquired between 2006 and 2019 and were collected for this analysis in compliance with the Health Insurance Portability and Accountability Act (HIPAA) and under institutional-review-board-approved protocols at each institution.

For each patient exam, the CC images of the left and right breast were used in analysis. Each patient included in this study had ultimately undergone core biopsy of an imaging abnormality with histopathologically confirmed findings of a malignant or benign lesion. However, it is important to note that all the images analyzed in this study were acquired prior to the detection of each mammographic abnormality (i.e., were antecedent images). The laterality of each mammographic abnormality was noted, and the affected and contralateral breasts were treated separately in the analyses.

The number of prior mammographic exams per participant ranged from 2 to 9 (Figure 1). Note that the period of time between subsequent screening exams was not always constant for each patient. The average time between exams was 1.27 years. The temporal mammograms for one patient, collected annually over a span of four years, are shown in Figure 2.

A total of 318 mammographic exams from 99 patients were included in the study. Of these, 49 patients were eventually diagnosed with a malignant finding and 50 were diagnosed with a benign finding. The mean age was 57.6 years (standard deviation 9.4 years) for the 49 cancer patients and 54.6 years (standard deviation 8.8 years) for the 50 cancer-free controls. All images were acquired on Hologic systems with pixel sizes of 70 μm × 70 μm and were processed according to the clinical standard at the patient's screening institution.

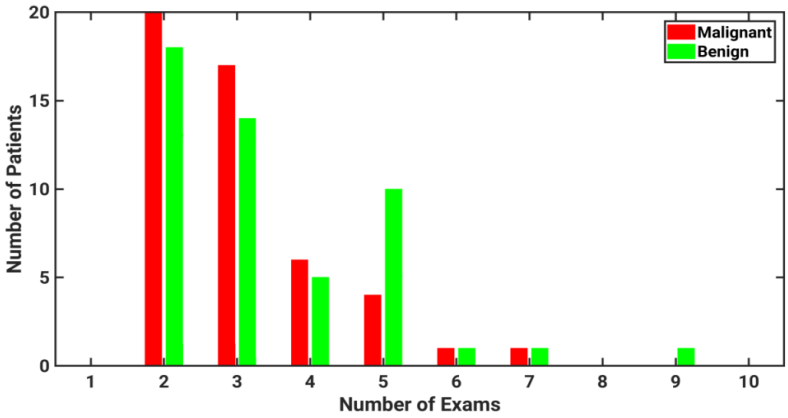

Figure 1. Histogram of the number of mammographic exams included in the study for patients with either malignant or benign lesions. All images included were acquired prior to the screening exam that ultimately led to diagnosis.

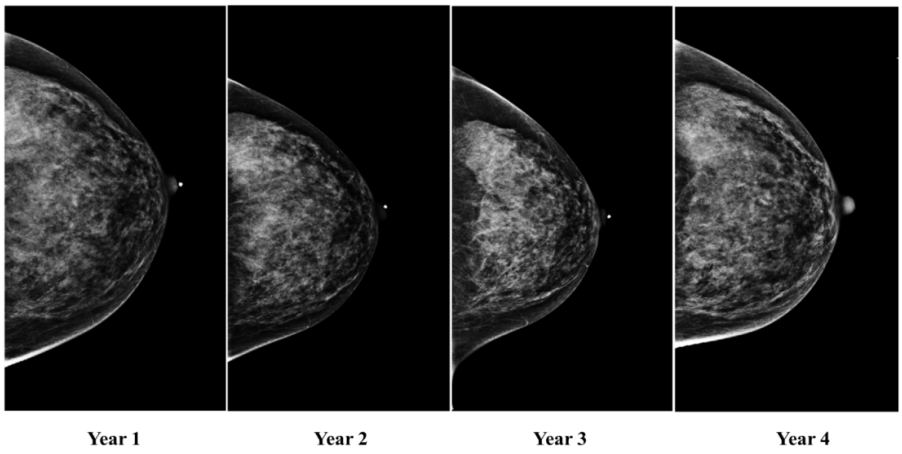

Figure 2. Temporal mammograms for one patient, collected annually over a span of four years.

2.2. Radiomic Feature Extraction

Computer-extracted radiomic features were automatically calculated on square ROIs of size 512 × 512 pixels, which had been manually placed in the central breast region posterior to the nipple. From each region, 50 features were automatically extracted and are summarized in Table 1. Additional details of these mathematical descriptors from feature extraction have been described elsewhere [11–16]. Features were selected to describe the intensity and spatial pattern of the texture in each image region.

Table 1. Summary of features included for analysis in the radiomics feature set.

Feature Category	Number of Features
Box counting fractal dimension	6
Edge gradient	4
Histogram	10
Fourier	2
Neighborhood Gray-Tone Difference Matrix	5
Minkowski fractal dimension	1
Powerlaw beta	8
GLCM	14
Total	50

2.3. Deep Feature Extraction

Deep-learning-based feature extraction was performed on the same ROIs used for radiomic feature calculation. Features were extracted using the pre-trained VGG-19 neural network [17]. A total of 1472 features were extracted from each image using the neural network. Features were extracted from each max pooling layer of the network, and an additional average pooling layer was added to reduce the dimensionality of the features. This approach of transfer learning has been studied and implemented elsewhere [18–20].

2.4. Long Short-Term Memory Network

Recurrent neural networks (RNNs) are designed for making classifications and predictions based on a time series of data [10]. RNNs are composed of a series of identical feed-forward neural networks. In this series of networks, each individual network is used to analyze a single time-point and is known as an RNN cell. Each RNN cell produces a recurrent output that is passed on to the next time step. Likewise, each RNN cell accepts a prior state as input. In this way, information from prior time-points informs the output of future time-points.

Mathematically, an RNN cell can be represented by Equation (1), where s_t is the current state, s_{t-1} is the prior state, x_t is the current input, and f is the recurrent function. Thus, a basic single-layer RNN can be written as in Equation (2), where ϕ is the activation function, and W, U, and b are the weights and biases of the network.

$$\begin{pmatrix} s_t \\ o_t \end{pmatrix} = f \begin{pmatrix} s_{t-1} \\ x_t \end{pmatrix} \tag{1}$$

$$s_t = \phi(W s_{t-1} + U x_t + b) \tag{2}$$

The general recurrent structure of an RNN is illustrated in Figure 3, where it is shown that information from the RNN cell for one time-point in the series is passed along to the cell for the next input from the series.

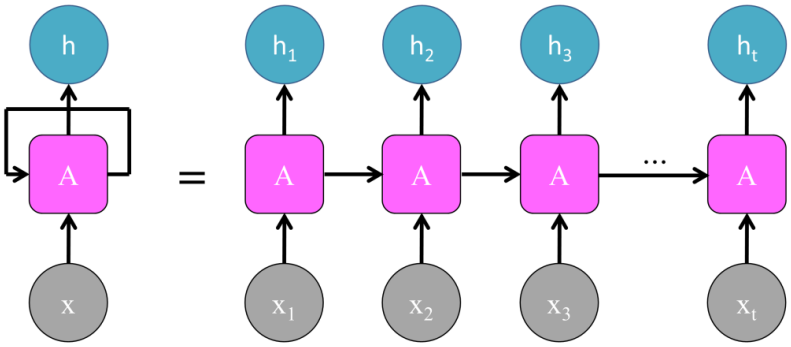

Figure 3. General architecture of an RNN cell component, where A represents the neural network, x_t represents the input, and h_t represents the output value [8].

In order to avoid the potential pitfalls of information morphing and of the vanishing gradient problem, LSTM cells are designed to contain three gates that are not typically present in conventional RNNs: the input gate, output gate, and forget gate. These three gates monitor the extent to which information is read in from an adjacent time-point, how much of this information to write out, and to what extent the information is remembered and passed on to the next time-point. The input gate (i_t), output gate (o_t), and forget gate (f_t) are defined as:

$$i_t = \sigma(W_i s_{t-1} + U_i x_t + b_i) \tag{3}$$

$$o_t = \sigma(W_o s_{t-1} + U_o x_t + b_o) \tag{4}$$

$$f_t = \sigma\left(W_f s_{t-1} + U_f x_t + b_f\right) \quad (5)$$

where s_{t-1} is the prior state, x_t is the current input, σ is the sigmoid function, and W, U, and b are the weights and biases of the network.

2.5. Classification and Evaluation

In order to evaluate the value of temporal information relative to single time-point analysis, classifications were performed using both SVM (single time-point) and LSTM (multiple time-points) in the task of predicting the histologic diagnosis of future lesions. The same feature set was used for training the LSTM and SVM networks. In this experiment, we decided to use SVM for comparisons as opposed to a feed-forward network in order to reduce the likelihood of overfitting. To characterize repeatability, 5-fold cross validation was used for each classifier, with folds kept consistent over each classifier along with the same proportions of malignant and benign cases in each fold. This method ensured that training and testing splits were kept consistent for pairwise comparisons between classifiers. Each classifier was trained separately on the antecedent images of the affected and contralateral breasts in the task of classifying the histologic diagnosis (cancer versus benign) of a future lesion. Note that images of any given case were kept together in either the training or testing fold.

ROI placement and radiomic feature extraction were performed on a dedicated workstation developed in our lab [12–15]. CNN feature extraction and network training were performed in Keras (Version 2.1.2) using a TensorFlow (Version 1.10.0) backend framework [21,22].

2.6. Temporal Sequence Classification with LSTM Network

In order to evaluate classification performance with the inclusion of multiple mammographic time-points, features extracted from each image were used as inputs to the LSTM network. To consider the value of the human-engineered radiomic features compared with the CNN features, separate networks were trained using each of these two as input features, as illustrated in Figure 4. Each classifier described was trained in the task of classifying future lesions as malignant or benign using only the prior antecedent images.

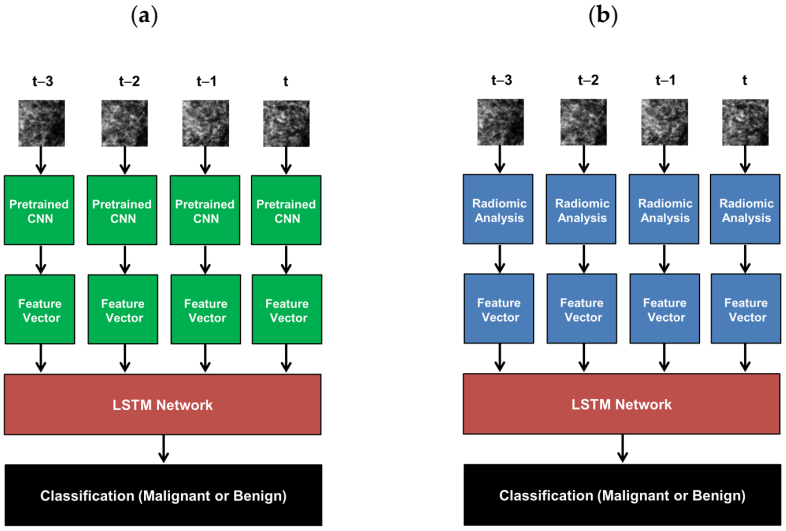

Figure 4. Summary of the workflow involved in using LSTM networks to classify temporal sequences of mammograms in this study, where t is the time of the most recent mammogram included in analysis. (**a**) Workflow for CNN-extracted features, and (**b**) workflow for radiomic features. Classifications were performed to predict the probabilities of future malignant lesions based only on antecedent images.

The LSTM network in this study was trained using a stochastic gradient descent (SGD) optimizer [23]. In SGD, optimal weights are determined by choosing a random sample of training vectors and using these to compute an estimate of the gradient at each step of the training procedure. Given a random batch of training objects, the update by SGD is given by Equation (6), where θ is the parameter to update, α is the learning rate, J is the objective function, and $(x^{(i)},y^{(i)})$ are the training feature vectors.

$$\theta = \theta - \alpha \nabla_\theta J\left(\theta; x^{(i)}, y^{(i)}\right) \quad (6)$$

Hyperparameters were selected by performing a limited sweep of learning rate and hidden dimension parameters. After sweeping over hidden dimensions of 512, 1024, and 2048, and sweeping over learning rates of 10^{-3}, 10^{-4}, and 10^{-5}, the combination of parameters of hidden dimension of 512 and learning rate of 10^{-4} was selected for the task of classifying future malignant lesions using antecedent images. Since each patient had a different number of images across the dataset, the feature sequences were padded with zeros to the length of the longest sequence, as typically conducted with LSTM. The padded part of the sequences was not taken into account when calculating the binary cross-entropy loss of the model [7]. For each LSTM network, 100 epochs were used in training.

2.7. Single Time-Point Classification with Support Vector Machine

To understand further the effect of multiple time-points, classification was also performed using the image collected only one year prior to diagnosis in the task of classifying the likelihood of malignancy of the future lesion. As only one single time-point is used for classification, a 5-fold cross validation using a support vector machine (SVM) as a classifier was performed [24]. To reduce dimensionality, principal component analysis (PCA) was performed to reduce the feature space to 25 principal components prior to training the SVM [25]. Training and classification were performed using the human-engineered radiomic features as an input, as well as using features extracted by the pretrained CNN as an input.

2.8. Statistical Evaluation

From receiver operating characteristic (ROC) analysis, the area under the curve (AUC) was used as a figure of merit in the task of predicting malignancy using antecedent images, and the statistical difference between the AUC values for different models was computed using ROCKIT software [26,27]. Corrections were made for multiple comparisons following the Holm–Bonferroni correction [28].

3. Results

The performance of each classification model for distinguishing a future benign state from a future malignant state is summarized in Table 2. In general, classifiers incorporating multiple time-points with LSTM, based either on deep-learning-extracted features or on radiomic features, tended to perform statistically better than chance (AUC = 0.5), whereas those using only a single time-point failed to show improved performance compared to chance, as judged by the area under the receiver operating characteristic curves (AUC: 0.63 ± 0.05, 0.65 ± 0.05, 0.52 ± 0.06 and 0.54 ± 0.06, respectively) for each of the affected breast and similarly for each of the contralateral breast (AUC: 0.59 ± 0.06, 0.62 ± 0.06, 0.52 ± 0.06 and 0.55 ± 0.06, respectively).

Note that we failed to show a significant difference in the AUC between the LSTM network trained using CNN-extracted features and that trained using radiomic features in the task of classifying future lesions as malignant or benign. This trend held for both classifications using the affected breast (AUCs of 0.63 vs. 0.65, p = 0.6511, 95% CI of ΔAUC [−0.1631, 0.1019]) and using the contralateral breast (AUCs of 0.59 vs. 0.62, p = 0.8083, 95% CI of ΔAUC [−0.1743, 0.1359]).

Table 2. Performance of each classification model for distinguishing future benign state from future malignant state.

Feature Type	LSTM Classifier AUC (*p*-Value) * [95% CI of AUC]	SVM Classifier AUC (*p*-Value) * [95% CI of AUC]
CNN (affected breast)	AUC = 0.63 (p = 0.0231) [0.5010, 0.7175]	AUC = 0.52 (p = 0.7103) [0.3962, 0.6193]
CNN (contralateral breast)	AUC = 0.59 (p = 0.1024) [0.4791, 0.6982]	AUC = 0.52 (p = 0.7389) [0.4083, 0.6320]
CNN (both lateralities)	AUC = 0.64 (p = 0.0104) [0.5184, 0.7336]	AUC = 0.54 (p = 0.5140) [0.4138, 0.6372]
Radiomics (affected breast)	AUC = 0.65 (p = 0.0042) [0.5346, 0.7456]	AUC = 0.54 (p = 0.4425) [0.4510, 0.6723]
Radiomics (contralateral breast)	AUC = 0.62 (p = 0.0259) [0.4998, 0.7161]	AUC = 0.55 (p = 0.3434) [0.4439, 0.6672]
Radiomics (both lateralities)	AUC = 0.63 (p = 0.0159) [0.5122, 0.7263]	AUC = 0.54 (p = 0.5035) [0.4216, 0.6454]
CNN + Radiomics (both lateralities)	AUC = 0.65 (p = 0.0059) [0.5109, 0.7282]	AUC = 0.52 (p = 0.7226) [0.4190, 0.6422]

* p-value is estimated using z-score test by comparing classifier performance with chance (AUC = 0.5). CI: confidence interval.

In clinical practice, it is unknown whether a future lesion will develop in the right or left breast. Therefore, it is more clinically relevant to examine a merged classifier, which takes into account the classifier output on each the left and right breast in the task of predicting whether the future lesion will be malignant or benign. We failed to demonstrate significant difference between classifiers trained using CNN features extracted from affected and contralateral breasts (AUCs of 0.63 vs. 0.59, p = 0.7278, 95% CI of ΔAUC [−0.0898, 0.1286]) and radiomic features extracted from affected and contralateral breasts (AUCs of 0.65 vs. 0.62, p = 0.6273, 95% CI of ΔAUC [−0.0211, 0.0350]).

Furthermore, it is also of interest to explore the classification performance in the task of characterizing future lesion malignancy when both the human-engineered and deep learning methods were combined over both breasts, as presented in Table 2 and Figure 5. Statistical comparisons were not performed on the merged classifier output in order to maintain statistical power by limiting the quantity of pairwise comparisons performed.

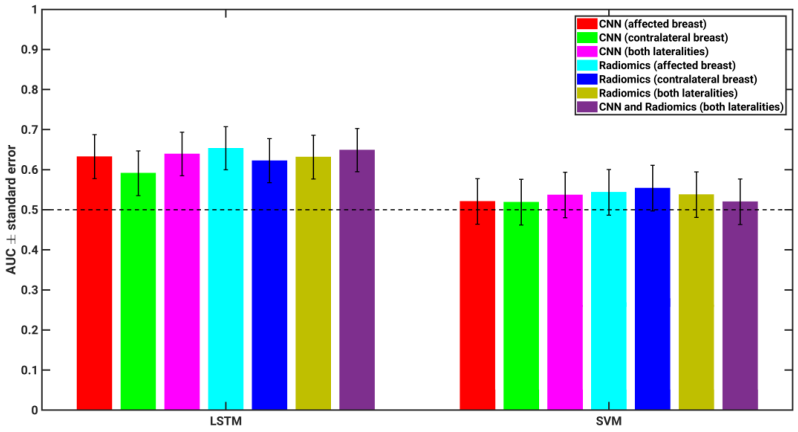

Figure 5. AUC values for each classifier compared, including merged classifiers. Each merged classifier was constructed by taking the average classifier output from two different classifiers for each individual case, and then performing ROC analysis on the averaged output values in the task of characterizing future lesions as malignant or benign. Error bars show one standard error. Dashed line is classification performance with guessing (AUC = 0.5).

4. Discussion

The results from this study demonstrated that LSTM classifiers using multiple time-points, based either on deep-learning-extracted features or on radiomic features, tended to perform statistically better than chance, whereas those using only a single time-point failed to show improved performance compared to chance, as judged by the area under the receiver operating characteristic curves (AUC: 0.63 ± 0.05, 0.65 ± 0.05, 0.52 ± 0.06 and 0.54 ± 0.06, respectively). The classification performance in the task of predicting future lesion malignancy was not observed to be statistically significantly different when an LSTM network was trained using either CNN features or using radiomic features. This suggests that, while these feature sets are different in their origin and how they are extracted, they achieve similar results. Thus, either feature set may be appropriate for classifications with temporal LSTM networks.

Similar classification performance was observed between the performance using features extracted from the affected and contralateral breast in predicting the malignancy of future unilateral lesions. Because only antecedent images were used in this analysis, no mammographic abnormalities were present. Thus, while it is possible that the affected breast had a precancerous texture change leading up to lesion detection, these results suggest that a change also occurred in the contralateral breast that may indicate future malignancy. Thus, this observation suggests that a field effect, changes in the breast extending beyond the lesion itself, is present in both the affected and contralateral breasts in antecedent imaging and, thus, the evaluation of either breast is informative for cancer risk assessment.

This investigation into the use of temporal sequences of data for malignancy prediction has several limitations. First, this study used a dataset of limited size compared with other implementations of LSTM networks. The curation of large datasets is more challenging and expensive in the medical domain compared with natural images, thus resulting in our small number of cases included. Additionally, the data used in this study were collected at two separate institutions, with slightly different cancer prevalence rates in the corresponding datasets. While all images were acquired on Hologic units, differences in image acquisition procedures may have varied between the two medical centers, resulting in some differences in image characteristics.

Additionally, the intervals at which women underwent screening were not consistent. While national agencies suggest screening at regular intervals of time, patient compliance was not consistent in the data. Furthermore, women may have undergone screening at an institution outside of the two involved in this study, and, therefore, this additional image was omitted from this investigation. Collecting images from consistent time intervals may affect, and potentially improve, the performance observed in this study.

The nature of screening exams involves repeat imaging on separate exam dates, thus inherently involving the repositioning of the patient in the imager. As a result of this, images are not spatially registered to one another. While this may be solved through deformable registration methods, it is likely that such image processing would alter the radiomic features extracted, potentially reducing the efficacy of such features. The approach taken in this study was to manually align ROIs on undeformed images; however, this method only results in approximate spatial registration across exam dates. While previous studies have shown that radiomic features tend to be only minimally impacted by small changes in spatial placement of an ROI, there may still be some effect present [29].

Finally, note that this study compared a new method, using LSTM networks to incorporate temporal information, with a conventional supervised learning approach (SVM) that does not involve deep learning. The transfer learning approach of using SVM to merge CNN-extracted image features has also shown promise in other FFDM studies [18,30].

This paper presents an image-based breast cancer prediction method that captures temporal information about parenchymal texture on FFDM over time. These temporal sequences are used to classify future lesions as either malignant or benign.

Compared with the previous methods, this work allowed for the incorporation of imaging information from multiple antecedent images, as opposed to just a single image. Thus, this method evaluated not only the appearance of the parenchyma, but also changes in the parenchyma over time. This work explored temporal network performance when using features extracted either by conventional radiomics methods and from the pre-trained VGG-19 network.

Based on the analyses performed in this study, LSTM networks based either on deep-learning-extracted features or on radiomic features from either affected breast or contralateral breast tended to perform statistically better than chance, whereas those using only a single time-point failed to show improved performance compared to chance.

The main motivation for the selection of LSTM networks for use in characterizing temporal image sequences is their ability to prevent vanishing or exploding gradients during error backpropagation. Additionally, LSTM networks are well suited to handle sequences of varying length, as women have varying numbers of screening mammograms throughout their lifetimes.

The method used in this study was motivated by the fact that human experts compare current screening mammograms with previous screening mammograms to assist in the detection of abnormality. This suggests that prior images may provide additional information to the current image [4]. Thus, changes in texture over time may be indicative of an elevated probability of developing a malignant breast lesion.

The deep learning methods employed here captured temporal data patterns that are not typically examined in conventional radiomics approaches. This work has shown that the temporal data patterns of either breast capture clinically useful information in evaluating the classification of future lesions based on screening mammography.

5. Conclusions

A long short-term memory (LSTM) network for the analysis of breast parenchyma, using both radiomic and deep-learning-based features, was performed to identify women predisposed to developing breast cancer. The findings from this study demonstrated that the incorporation of temporal information into radiomic analyses may improve overall classification performance through LSTM, as demonstrated by the improved discrimination of future lesions as malignant or benign. Further, our data suggest that a potential field effect, changes in the breast extending beyond the lesion itself, is present in both the affected and contralateral breasts in antecedent imaging, and, thus, the evaluation of either breast might inform on the future risk of breast cancer.

Author Contributions: Conceptualization, H.L., K.R., I.B. and M.L.G.; methodology, H.L., K.R. and M.L.G.; software, H.L., K.R., L.L. and C.-W.C.; validation, H.L.; formal analysis, H.L.; investigation, H.L., K.R., I.B. and M.L.G.; resources, H.L., L.L., N.B., C.-W.C., M.E., G.J.W., R.E.-Z., I.B. and M.L.G.; data curation, H.L., K.R., L.L., N.B., M.E., G.J.W., R.E.-Z. and I.B.; writing—original draft preparation, H.L. and K.R.; writing—review and editing, H.L., K.R., L.L., N.B., C.-W.C., M.E., G.J.W., R.E.-Z., I.B. and M.L.G.; visualization, H.L.; supervision, I.B. and M.L.G.; project administration, H.L.; funding acquisition, R.E.-Z., I.B. and M.L.G. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by NIH 1U01CA189240-01 grant to I.B. and El-Zein (M-PI), and KG090351 grant from the Susan G. Komen Foundation to I.B.; NIH U01CA195564, T32 EB002103, S10 OD025081 and University of Chicago Comprehensive Cancer Center Koleseiki Funding to M.L.G.; NIH F31CA228247 grant to K.R.

Institutional Review Board Statement: The database for this study was retrospectively collected under Health Insurance Portability and Accountability Act (HIPAA)-compliant Institutional Review Board (IRB) protocols.

Informed Consent Statement: All clinical information and images in this study were deidentified to the investigators, and, hence, the IRB of the University of Chicago and The University of Texas MD Anderson Cancer Center approved the study and waived informed consent from the participants.

Data Availability Statement: Radiomic data are available upon reasonable request to the authors.

Conflicts of Interest: M.L.G. is a stockholder in R2 technology/Hologic and QView, receives royalties from UChicago Tech, and is a cofounder in Quantitative Insights (now Qlarity Imaging). H.L. and L.L. receive royalties from UChicago Tech. It is the University of Chicago Conflict of Interest Policy that investigators disclose publicly actual or potential significant financial interest that would reasonably appear to be directly and significantly affected by the research activities. The other authors have no relevant conflicts to disclose.

References

1. Smith, R.A.; Cokkinides, V.; Brawley, O.W. Cancer screening in the United States, 2009: A review of current American Cancer Society guidelines and issues in cancer screening. *CA Cancer J. Clin.* **2009**, *59*, 27–41. [CrossRef]
2. Qaseem, A.; Snow, V.; Sherif, K.; Aronson, M.; Weiss, K.B.; Owens, D.K.; Clinical Efficacy Assessment Subcommittee of the American College of Physicians. Screening Mammography for Women 40 to 49 Years of Age: A Clinical Practice Guideline from the American College of Physicians. *Ann. Intern. Med.* **2007**, *146*, 511–515. [CrossRef] [PubMed]
3. Oeffinger, K.C.; Fontham, E.T.H.; Etzioni, R.; Herzig, A.; Michaelson, J.S.; Shih, Y.-C.T.; Walter, L.C.; Church, T.R.; Flowers, C.R.; LaMonte, S.J.; et al. Breast Cancer Screening for Women at Average Risk. *JAMA* **2015**, *314*, 1599–1614. [CrossRef] [PubMed]
4. Yankaskas, B.C.; May, R.C.; Matuszewski, J.; Bowling, J.M.; Jarman, M.P.; Schroeder, B.F. Effect of Observing Change from Comparison Mammograms on Performance of Screening Mammography in a Large Community-based Population. *Radiology* **2011**, *261*, 762–770. [CrossRef] [PubMed]
5. Santeramo, R.; Withey, S.; Montana, G. Longitudinal Detection of Radiological Abnormalities with Time-Modulated LSTM. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 326–333.
6. Shao, Y.; Chen, Z.; Ming, S.; Ye, Q.; Shu, Z.; Gong, C.; Pang, P.; Gong, X. Predicting the Development of Normal-Appearing White Matter With Radiomics in the Aging Brain: A Longitudinal Clinical Study. *Front. Aging Neurosci.* **2018**, *10*, 393. [CrossRef] [PubMed]
7. Antropova, N.; Huynh, B.; Li, H.; Giger, M.L. Breast lesion classification based on dynamic contrast-enhanced magnetic resonance images sequences with long short-term memory networks. *J. Med. Imaging* **2018**, *6*, 011002. [CrossRef]
8. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
9. Hochreiter, S. The Vanishing Gradient Problem During Learning Recurrent Neural Nets and Problem Solutions. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.* **1998**, *6*, 107–116. [CrossRef]
10. Zaremba, W.; Sutskever, I.; Vinyals, O. Recurrent Neural Network Regularization. *arXiv* **2014**, arXiv:1409.2329.
11. Mendel, K.R.; Li, H.; Lan, L.; Cahill, C.M.; Rael, V.; Abe, H.; Giger, M.L. Quantitative texture analysis: Robustness of radiomics across two digital mammography manufacturers' systems. *J. Med. Imaging* **2018**, *5*, 11002. [CrossRef]
12. Huo, Z.; Giger, M.L.; Wolverton, D.E.; Zhong, W.; Cumming, S.; Olopade, O.I. Computerized analysis of mammographic parenchymal patterns for breast cancer risk assessment: Feature selection. *Med. Phys.* **2000**, *27*, 4–12. [CrossRef] [PubMed]
13. Li, H.; Giger, M.L.; Olopade, O.I.; Lan, L. Fractal Analysis of Mammographic Parenchymal Patterns in Breast Cancer Risk Assessment. *Acad. Radiol.* **2007**, *14*, 513–521. [CrossRef] [PubMed]
14. Li, H.; Giger, M.L.; Olopade, O.I.; Chinander, M.R. Power Spectral Analysis of Mammographic Parenchymal Patterns for Breast Cancer Risk Assessment. *J. Digit. Imaging* **2008**, *21*, 145–152. [CrossRef] [PubMed]
15. Li, H.; Giger, M.L.; Lan, L.; Janardanan, J.; Sennett, C.A. Comparative analysis of image-based phenotypes of mammographic density and parenchymal patterns in distinguishing between *BRCA1/2* cases, unilateral cancer cases, and controls. *J. Med. Imaging* **2014**, *1*, 031009. [CrossRef] [PubMed]
16. Haralick, R.M.; Shanmugam, K.; Dinstein, I.H. Textural Features for Image Classification. *IEEE Trans. Syst. Man Cybern.* **1973**, *6*, 610–621. [CrossRef]
17. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2017**, arXiv:1409.1556.
18. Huynh, B.Q.; Li, H.; Giger, M.L. Digital mammographic tumor classification using transfer learning from deep convolutional neural networks. *J. Med. Imaging* **2016**, *3*, 034501. [CrossRef]
19. Mendel, K.; Li, H.; Sheth, D.; Giger, M. Transfer Learning from Convolutional Neural Networks for Computer-Aided Diagnosis: A Comparison of Digital Breast Tomosynthesis and Full-Field Digital Mammography. *Acad. Radiol.* **2019**, *26*, 735–743. [CrossRef]
20. Antropova, N.; Huynh, B.Q.; Giger, M.L. A deep feature fusion methodology for breast cancer diagnosis demonstrated on three imaging modality datasets. *Med. Phys.* **2017**, *44*, 5162–5171. [CrossRef]
21. Abadi, M.; Barham, P.; Chen, J.; Chen, Z.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Irving, G.; Isard, M.; et al. TensorFlow: A System for Large-Scale Machine Learning. In Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI '16), Savannah, GA, USA, 2–4 November 2016.
22. Gulli, A.; Pal, S. *Deep Learning with Keras*; Packt Publishing Ltd.: Birmingham, UK, 2017.
23. Zhang, T. Solving Large Scale Linear Prediction Problems Using Stochastic Gradient Descent Algorithms. In Proceedings of the Twenty-First International Conference on Machine Learning, Banff, AB, Canada, 4–8 July 2004; p. 116. [CrossRef]

24. Suykens, J.A.K.; Vandewalle, J. Least Squares Support Vector Machine Classifiers. *Neural Process. Lett.* **1999**, *9*, 293–300. [CrossRef]
25. Wold, S.; Esbensen, K.; Geladi, P. Principal component analysis. *Chemom. Intell. Lab. Syst.* **1987**, *2*, 37–52. [CrossRef]
26. Metz, C.E.; Herman, B.A.; Roe, C.A. Statistical Comparison of Two ROC-curve Estimates Obtained from Partially-paired Datasets. *Med. Decis. Mak.* **1998**, *18*, 110–121. [CrossRef] [PubMed]
27. Metz, C.E. Basic Principles of ROC Analysis. In *Seminars in Nuclear Medicine*; WB Saunders: Philadelphia, PA, USA, 2018; Available online: http://gim.unmc.edu/dxtests/ROC1.htm (accessed on 4 May 2018).
28. Holm, S. A Simple Sequentially Rejective Multiple Test Procedure. *Scand. J. Stat.* **1979**, *6*, 65–70.
29. Gierach, G.L.; Li, H.; Loud, J.T.; Greene, M.H.; Chow, C.K.; Lan, L.; Prindiville, S.A.; Eng-Wong, J.; Soballe, P.W.; Giambartolomei, C.; et al. Relationships between computer-extracted mammographic texture pattern features and BRCA1/2 mutation status: A cross-sectional study. *Breast Cancer Res.* **2014**, *16*, 424. [CrossRef] [PubMed]
30. Li, H.; Giger, M.L.; Huynh, B.Q.; Antropova, N.O. Deep learning in breast cancer risk assessment: Evaluation of convolutional neural networks on a clinical dataset of full-field digital mammograms. *J. Med Imaging* **2017**, *4*, 041304. [CrossRef]