Title

Prospective Analysis Using a Novel CNN Algorithm to Distinguish Atypical Ductal Hyperplasia From Ductal Carcinoma in Situ in Breast.

Journal

Clinical Breast Cancer, 20(6)

Authors

Mutasa, Simukayi

Chang, Peter

Nemer, John

et al.

Publication Date

2020-12-01

DOI

10.1016/j.clbc.2020.06.001

Peer reviewed

HHS Public Access

Author manuscript

Clin Breast Cancer. Author manuscript; available in PMC 2021 June 16.

Published in final edited form as:

Clin Breast Cancer. 2020 December ; 20(6): e757–e760. doi:10.1016/j.clbc.2020.06.001.

Prospective Analysis Using a Novel CNN Algorithm to Distinguish Atypical Ductal Hyperplasia From Ductal Carcinoma in Situ in Breast

Simukayi Mutasa[1], **Peter Chang**[2], **John Nemer**[1], **Eduardo Pascual Van Sant**[1], **Mary Sun**[1], **Alison McIlvride**[1], **Maham Siddique**[1], **Richard Ha**[1,3]

[1]Department of Radiology, Columbia University Medical Center, New York, NY

[2]Center for Artificial Intelligence in Diagnostic Medicine (CAIDM), Division of Neuroradiology, UCI Health, Department of Radiological Sciences, Orange, CA

[3]Breast Imaging Section, Columbia University Medical Center, New York, NY

Abstract

The purpose of this study is to prospectively validate our previously developed convolutional neural networks algorithm using 280 unseen mammographic images to distinguish between pure atypical ductal hyperplasia from ductal carcinoma in situ. With a specificity of 93.7%, it is feasible to use our convolutional neural networks algorithm to identify patients with pure atypical ductal hyperplasia who may be safely observed rather than undergo surgery.

Introduction: We previously developed a convolutional neural networks (CNN)-based algorithm to distinguish atypical ductal hyperplasia (ADH) from ductal carcinoma in situ (DCIS) using a mammographic dataset. The purpose of this study is to further validate our CNN algorithm by prospectively analyzing an unseen new dataset to evaluate the diagnostic performance of our algorithm.

Materials and Methods: In this institutional review board-approved study, a new dataset composed of 280 unique mammographic images from 140 patients was used to test our CNN algorithm. All patients underwent stereotactic-guided biopsy of calcifications and underwent surgical excision with available final pathology. The ADH group consisted of 122 images from 61 patients with the highest pathology diagnosis of ADH. The DCIS group consisted of 158 images from 79 patients with the highest pathology diagnosis of DCIS. Two standard mammographic magnification views (craniocaudal and mediolateral/lateromedial) of the calcifications were used for analysis. Calcifications were segmented using an open source software platform 3D slicer and resized to fita 128 128 pixel bounding box. Our previously developed CNN algorithm was used. Briefly, a 15 hidden layer topology was used. The network architecture contained 5 residual layers and dropout of 0.25 after each convolution. Diagnostic performance metrics were analyzed including sensitivity, specificity, accuracy, and area under the receiver operating characteristic

Address for correspondence: Richard Ha, MD, MS, Associate Professor of Radiology, Director of Research and Education, Breast Imaging Section, Columbia University Medical Center, 622 W 168th St, PB-1-301, New York, NY 10032 rh2616@cumc.columbia.edu.

Disclosure

The authors have stated that they have no conflicts of interest.

curve. The “positive class” was defined as the pure ADH group in this study and thus specificity represents minimizing the amount of falsely labeled pure ADH cases.

Results: Area under the receiver operating characteristic curve was 0.90 (95% confidence interval, ± 0.04). Diagnostic accuracy, sensitivity, and specificity was 80.7%, 63.9%, and 93.7%, respectively.

Conclusion: Prospectively tested on new unseen data, our CNN algorithm distinguished pure ADH from DCIS using mammographic images with high specificity.

Keywords

ADH; Artificial intelligence; Convolutional neural networks; DCIS; Deep learning

Introduction

Breast cancer is the most common cancer affecting women worldwide and the second most common cause of cancer deaths among women in the United States.[1] Atypical ductal hyperplasia (ADH) is a non-obligate precursor to invasive disease, and is characterized by high-risk proliferation of epithelial cells in the terminal ductal lobular units of the breast.[2] ADH is diagnosed in up to 15% of biopsies following suspicious screen-detected lesions. Surgical excision is the current standard of care for ADH, with upgrade rates to ductal carcinoma in situ (DCIS) or invasive cancer of 10% to 30% at the time of excision.[3] The majority of patients undergo surgical excision without upgrade to malignancy, indicating a need to identify patients who may be more appropriate for observation and potentially spare the cost and morbidity associated with surgery.

Previously, several groups have attempted to identify a favorable subset of low-risk patients with a diagnosis of ADH who may be observed rather than undergo surgery based on various clinical, histologic, and/or radiographic criteria, with limited success.[3–9] Unfortunately, these retrospective studies have not thus far changed management recommendations. More recently, we developed a convolutional neural network (CNN)-based algorithm to distinguish ADH from DCIS using a mammographic dataset, yielding a high degree of diagnostic accuracy (86.7%) that has potential for clinical application.[10]

An artificial neural network, such as a CNN, is a computational model trained to recognize image features through the extraction of abstract features from image datasets, and these are currently being used with medical imagining for classification tasks with success on mostly retrospective feasibility studies.[5] Despite great enthusiasm and promise, only a small subset of research studies utilizing artificial intelligence (AI) and CNN in diagnostic analysis have demonstrated clinical performance or validation.[11] The use of new datasets is especially important in preventing overestimation of results generated by the algorithm by reducing overfitting and spectrum bias.[12] In addition, the prospective analysis of an unseen dataset provides much stronger evidence for clinical efficacy than retrospective case-control studies in predicting the real-world outcome.[13]

The purpose of this study is to use our CNN-based algorithm to prospectively distinguish ADH from DCIS using a new unseen mammographic dataset.

Materials and Methods

This study was approved by our institutional review board. A total of 280 unique mammographic images from 140 patients who underwent consecutive stereotactic-guided biopsies, seen at our institution from January 2016 to February 2018, were used to test our CNN algorithm. All patients presented with suspicious calcifications with no associated mass on mammography. All patients had atleast 2 magnification views, a craniocaudal view and either a mediolateral or lateromedial view. Mammography was performed using dedicated mammography units (Senographe Essential, GE Healthcare). All patients underwent subsequent surgical excision with final available pathology.

Clinical and pathologic data were collected, including patient age, span of calcifications, and pathologic result. Standard pathology guideline was used for interpretation.[2] Briefly, the presence of DCIS was confirmed on the basis of nuclear grade, necrosis, cell polarization, and architectural pattern. The criteria used to distinguish ADH from DCIS included the presence of at least 1 of 2 quantitative features: size limited to 2 mm or smaller and involvement of no more than 2 membrane-bound spaces.[2]

Descriptive statistics were used to summarize clinical, imaging, and pathologic parameters. We performed a 2-sample *t* test for each of these variables on the basis of normal distribution. All statistical analyses were performed using a statistical software program (SPSS Statistics for Microsoft Windows, version 24, SPSS). A 2-sided $P \leq .05$ was considered significant.

Data Segmentation, Augmentation, and CNN Architecture

Previously described CNN methodology was used.[10] Briefly, mammographic images were loaded into a 3D segmentation program. A fellowship-trained breast radiologist (R.H.) with 10 years of experience manually extracted segmentations to encompass the regions of the magnification view that contained calcifications. Each image was scaled in size on the basis of the radius of the segmentations and was resized to fit a bounding box of 128 × 128 pixels. The entire image batch was centered using histogram-based z score normalization of the non-air pixel intensity values. Augmentation was performed by the following: Images were randomly flipped vertically, horizontally, or in both directions; were rotated by a random angle between 0.52 and −0.52 radians; and were randomly cropped to a box 80% of the initial size. In addition, random affine shear was applied to each input breast image. A CNN topology with 15 hidden layers was used to implement the neural network (Figure 1). The network architecture contained 5 residual layers and dropout of 0.25 after each convolution. Software code for this study was written using an open-source software library for numerical computation (Python TensorFlow library, version 1.5). Experiments and network training were performed on a workstation with an Ubuntu operating system (release 16.04, Canonical) and a graphics card (Titan X Pascal, NVIDIA).

Diagnostic performance metrics were analyzed, including sensitivity, specificity, accuracy, and area under the receiver operating characteristic curve (AUC). The “positive class” was defined as the pure ADH group in this study, and thus specificity represents minimizing the amount of falsely labeled pure ADH cases.

Results

Of 140 patients, 61 patients yielded the highest pathology diagnosis of ADH on biopsy and on subsequent surgical excision. Of 140 patients, 79 patients yielded the highest pathology diagnosis of DCIS on final surgical excision. The mean patient's age with ADH diagnosis was 56.1 ± 12.2 years, and the mean patient's age with DCIS was 61.9 ± 11.1 years. The difference in age between the 2 groups was statistically significant ($P = .01$). The mean distance of mammographic calcifications' extent was 1.09 ± 0.95 cm in the ADH group and 1.37 ± 1.0 cm in the DCIS group. The difference in the extent of the mammographic calcifications between the 2 groups was not significant ($P = .10$). The mean number of core samples obtained per biopsy was 8.8 (± 2.7) cores for the patients with ADH and 8.9 (± 2.6) cores for the patients with DCIS. The number of cores between the 2 groups was not significantly different ($P = .24$). For the 79 patients with a DCIS diagnosis, the grade was determined to be low or intermediate for 46 patients and high for 33 patients.

A total of 280 unique images representing mediolateral and craniocaudal magnification views of calcifications were used for CNN analysis; 122 images from 61 patients with the final diagnosis of ADH and 158 images from 79 patients with the final diagnosis of DCIS. CNN analysis of distinguishing ADH from DCIS yielded an AUC of 0.90 (95% confidence interval [CI], ± 0.04) (Figure 2). Diagnostic accuracy, sensitivity, and specificity was 80.7%, 63.9%, and 93.7%, respectively.

Discussion

Using our CNN algorithm, we prospectively analyzed unseen data to distinguish ADH from DCIS, yielding a diagnostic accuracy of 80.7%. This could potentially be used in the clinical setting as a valuable decision-making support tool to determine which patients with an initial diagnosis of ADH on core biopsy may be safely observed rather than require further surgical excision.

Despite great excitement regarding AI technology, many published studies using AI tools and medical imaging are single-institution, retrospective, feasibility studies. A study by Kim et al, evaluating published AI studies, concluded that the majority of them lacked the study design necessary for clinical validation.[11] The use of our CNN algorithm on an unseen dataset analyzed in a prospective manner was purposely designed for further clinical validation. This approach limits overfitting, in which a CNN algorithm becomes overly reliant on the provided training data, which is detrimental to the generalization of new data while simultaneously artificially boosting model performance.[12] Overfitting is a frequently encountered challenge in AI algorithm development. Our study of testing a CNN algorithm in a new unseen dataset is an important step in transitioning this new technology out of the research lab and into the clinic.

Our CNN algorithm was designed to maximize diagnostic specificity in order to minimize DCIS cases miscategorized as pure ADH cases, yielding a high specificity of 93.7% while maintaining a reasonable diagnostic accuracy of 80.7% and an AUC of 0.90. High specificity will enable careful selection of patients that may potentially be observed rather

than require surgery. Previous attempts by several groups to identify low-risk patients who may be observed with a diagnosis of ADH rather than undergo surgery based on various clinical, histologic, and/or radiographic criteria yielded mostly inferior diagnostic performance compared with our study.[3–9] In addition, the implementation of some of these prediction models requires criteria that are not always feasible to achieve, such as removal of 95% of calcifications. Other studies using breast magnetic resonance imaging (MRI) data have shown potential but with some limitations, including small sample sizes owing to patients diagnosed with atypia who do not routinely undergo breast MRI.[14,15] Another major limitation of using breast MRI data is that the breast MRI is usually performed after the diagnosis by core needle biopsy, which results in a significant amount of tissue removal, thus limiting the interpretive value of the remaining tissue. Thus far, these types of studies have variable results with minimal consensus on which patients can be safely selected to undergo observation after a diagnosis of ADH.

Although histopathologic analysis is the current gold standard for distinguishing ADH from DCIS, it is also subjective, prone to interobserver variability, and limited by the amount of tissue obtained.[16] The distinct advantage of applying a CNN model to pre-biopsy mammogram images is that these images capture the region of interest prior to biopsy, allowing for a comprehensive analysis of the whole region of interest. In addition, once the region of interest is identified manually, the interpretation of the region by the CNN can be fully automated, which will enable rapid and objective evaluation of the mammographic images without inter- and intra-observer variability.

Similar to the results of our previous study,[10] patients with ADH were significantly younger than patients with DCIS. Although the clinical significance of this finding is unclear, a possible reason includes the fact that ADH is considered a potential precursor to DCIS and thus may occur more commonly in younger women. In addition, a large study using the Breast Cancer Surveillance Consortium showed higher rates of ADH in younger patients compared with higher rates of ADH and cancer in older patients.[17]

Although our study design to test our CNN algorithm on unseen new data will be helpful in the process of clinical validation, the dataset is still relatively small and from a single institution. Further validation will be needed with data from multiple institutions performed prospectively with randomization as described by Kim et al.[11] In addition, further training with a larger dataset will likely improve our model. The performance of CNN has been shown to increase logarithmically with larger datasets.[18] Furthermore, the potential of combining clinical information and the results of our CNN algorithm in order to further improve the overall prediction model is under investigation. Lastly, because training a CNN is an end-to-end process, it does not clearly identify the reasoning in a deterministic manner. This is an ongoing area of research to improve human understanding and intuition behind the predictions of an artificial neural network.

Our CNN algorithm was able to distinguish ADH from DCIS with a high degree of specificity (93.7%) using a new unseen mammographic dataset. This can potentially aid in appropriate patient selection for observation in patients diagnosed with ADH on core biopsy rather than surgery.

Acknowledgments

NVIDIA GPU provided by the GPU Grant Program, NVIDIA Corporation.

References

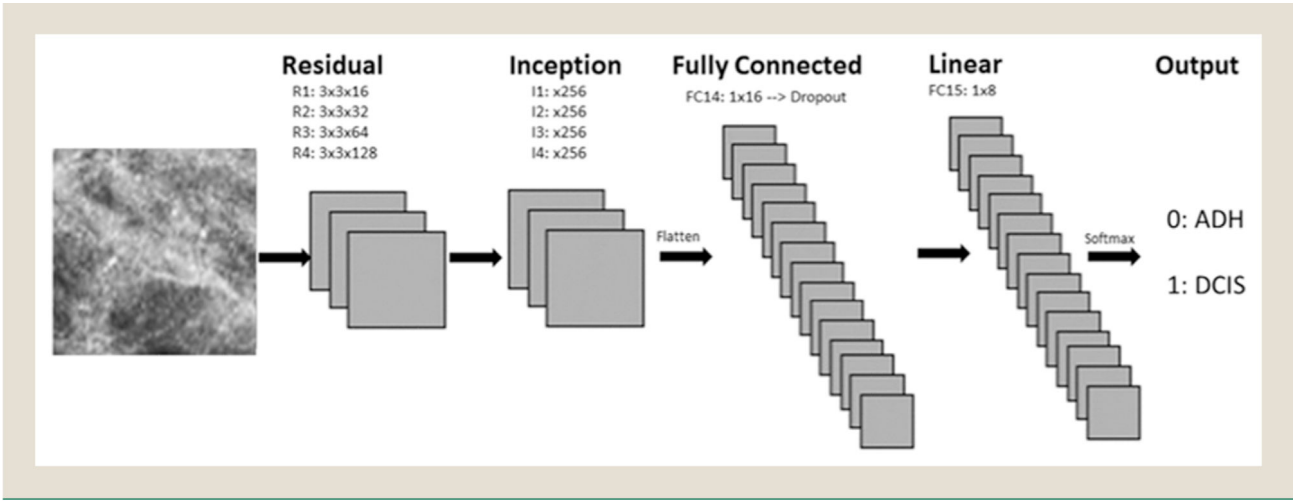
1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2019. CA Cancer J Clin 2019; 69: 7–34. [PubMed: 30620402]
2. Sinn HP, Kreipe H. A Brief overview of the WHO Classification of Breast Tumors, 4th Edition, focusing on issues and updates from the 3rd edition. Breast Care (Basel) 2013; 8:149–54. [PubMed: 24415964]
3. Menen RS, Ganesan N, Bevers T, et al. Long-term safety of observation in selected women following core biopsy diagnosis of atypical ductal hyperplasia. Ann Surg Oncol 2017; 24:70–6. [PubMed: 27573525]
4. Bendifallah S, Defert S, Chabbert-Buffet N, et al. Scoring to predict the possibility of upgrades to malignancy in atypical ductal hyperplasia diagnosed by an 11-gauge vacuum-assisted biopsy device: an external validation study. Eur J Cancer 2012; 48: 30–6. [PubMed: 22100905]
5. Chen LY, Hu J, Tsang JYS, et al. Diagnostic upgrade of atypical ductal hyperplasia of the breast based on evaluation of histopathological features and calcification on core needle biopsy. Histopathology 2019; 75:320–8. [PubMed: 31013355]
6. Deshaies I, Provencher L, Jacob S, et al. Factors associated with upgrading to malignancy at surgery of atypical ductal hyperplasia diagnosed on core biopsy. Breast 2011; 20:50–5. [PubMed: 20619647]
7. Ko E, Han W, Lee JW, et al. Scoring system for predicting malignancy in patients diagnosed with atypical ductal hyperplasia at ultrasound-guided core needle biopsy. Breast Cancer Res Treat 2008; 112:189–95. [PubMed: 18060577]
8. Nguyen CV, Albarracin CT, Whitman GJ, Lopez A, Sneige N. Atypical ductal hyperplasia in directional vacuum-assisted biopsy of breast microcalcifications: considerations for surgical excision. Ann Surg Oncol 2011; 18:752–61. [PubMed: 20972636]
9. Pankratz VS, Hartmann LC, Degnim AC, et al. Assessment of the accuracy of the Gail model in women with atypical hyperplasia. J Clin Oncol 2008; 26:5374–9. [PubMed: 18854574]
10. Ha R, Mutasa S, Sant EPV, et al. Accuracy of distinguishing atypical ductal hyperplasia from ductal carcinoma in situ with convolutional neural network-based machine learning approach using mammographic image data. AJR Am J Roent-genol 2019; 5:1166–71.
11. Kim DW, Jang HY, Kim KW, Shin Y, Park SH. Design characteristics of studies reporting the performance of artificial intelligence algorithms for diagnostic analysis of medical images: results from recently published papers. Korean J Radiol 2019; 20:405–10. [PubMed: 30799571]
12. Park SH, Han K. Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. Radiology 2018; 286:800–9. [PubMed: 29309734]
13. Park SH. Diagnostic case-control versus diagnostic cohort studies for clinical validation of artificial intelligence algorithm performance. Radiology 2019; 290: 272–3. [PubMed: 30511912]
14. Amitai Y, Menes T, Golan O. Use of breast magnetic resonance imaging in women diagnosed with atypical ductal hyperplasia at core needle biopsy helps select women for surgical excision. Can Assoc Radiol J 2018; 69:240–7. [PubMed: 29958833]
15. Tsuchiya K, Mori N, Schacht DV, et al. Value of breast MRI for patients with a biopsy showing atypical ductal hyperplasia (ADH). J Magn Reson Imaging 2017; 46:1738–47. [PubMed: 28295791]
16. Gomes DS, Porto SS, Balabram D, Gobbi H. Inter-observer variability between general pathologists and a specialist in breast pathology in the diagnosis of lobular neoplasia, columnar cell lesions, atypical ductal hyperplasia and ductal carcinoma in situ of the breast. Diagn Pathol 2014; 9:121. [PubMed: 24948027]
17. Menes TS, Kerlikowske K, Jaffer S, Seger D, Miglioretti DL. Rates of atypical ductal hyperplasia have declined with less use of postmenopausal hormone treatment: findings from the Breast

Cancer Surveillance Consortium. Cancer Epidemiol Biomarkers Prev 2009; 18:2822–8. [PubMed: 19900937]

18. Truhn D, Schrading S, Haarburger C, Schneider H, Merhof D, Kuhl C. Radiomic versus convolutional neural networks analysis for classification of contrast-enhancing lesions at multiparametric breast MRI. Radiology 2019; 290:290–7. [PubMed: 30422086]

Clinical Practice Points

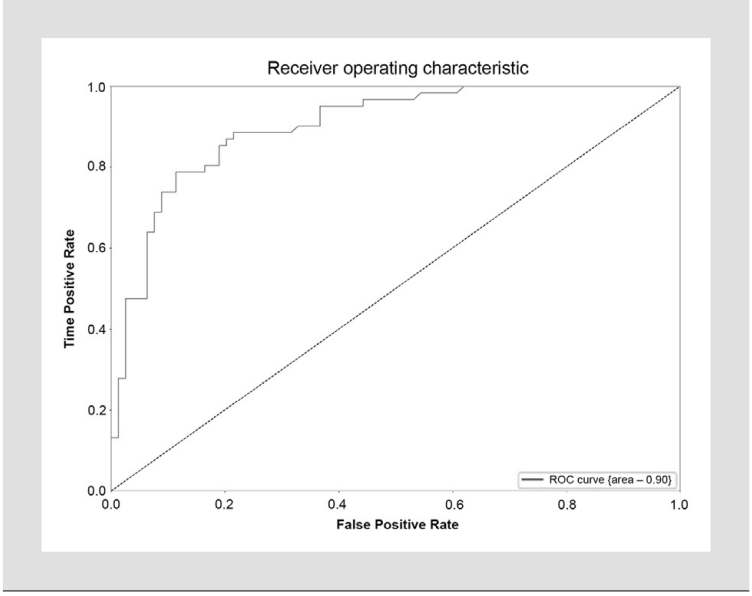
- Using the patients' mammographic images, our CNN algorithm can be used to predict patients with pure ADH who may be safely observed rather than undergo surgery.



Abbreviations: ADH = atypical ductal hyperplasia; DCIS = ductal carcinoma in situ.

Figure 1.

The Convolutional Neural Network Architecture for 2-Class Prediction of ADH Versus DCIS



Abbreviations: ADH = atypical ductal Hyperplasia; DCIS = ductal carcinoma in situ; ROC = receiver operating characteristic.

Figure 2.

Area Under the ROC Curve of 2-Classification Convolutional Neural Network Model (ADH vs. DCIS)