ORIGINAL ARTICLE

Prediction of Occult Invasive Disease in Ductal Carcinoma *in Situ* Using Deep Learning Features

Bibo Shi, PhD[a], Lars J. Grimm, MD[a], Maciej A. Mazurowski, PhD[a], Jay A. Baker, MD[a], Jeffrey R. Marks, PhD[b], Lorraine M. King, PhD[b], Carlo C. Maley, PhD[c,d], E. Shelley Hwang, MD[b], Joseph Y. Lo, PhD[a]

Abstract

Purpose: The aim of this study was to determine whether deep features extracted from digital mammograms using a pretrained deep convolutional neural network are prognostic of occult invasive disease for patients with ductal carcinoma in situ (DCIS) on core needle biopsy.

Methods: In this retrospective study, digital mammographic magnification views were collected for 99 subjects with DCIS at biopsy, 25 of which were subsequently upstaged to invasive cancer. A deep convolutional neural network model that was pretrained on nonmedical images (eg, animals, plants, instruments) was used as the feature extractor. Through a statistical pooling strategy, deep features were extracted at different levels of convolutional layers from the lesion areas, without sacrificing the original resolution or distorting the underlying topology. A multivariate classifier was then trained to predict which tumors contain occult invasive disease. This was compared with the performance of traditional "handcrafted" computer vision (CV) features previously developed specifically to assess mammographic calcifications. The generalization performance was assessed using Monte Carlo cross-validation and receiver operating characteristic curve analysis.

Results: Deep features were able to distinguish DCIS with occult invasion from pure DCIS, with an area under the receiver operating characteristic curve of 0.70 (95% confidence interval, 0.68-0.73). This performance was comparable with the handcrafted CV features (area under the curve = 0.68; 95% confidence interval, 0.66-0.71) that were designed with prior domain knowledge.

Conclusions: Despite being pretrained on only nonmedical images, the deep features extracted from digital mammograms demonstrated comparable performance with handcrafted CV features for the challenging task of predicting DCIS upstaging.

Key Words: Ductal carcinoma in situ, digital mammography, convolution neural network, deep learning, computer vision

J Am Coll Radiol 2017;■:■-■. Copyright © 2017 American College of Radiology

INTRODUCTION

More than 60,000 women in the United States are diagnosed with ductal carcinoma in situ (DCIS) every year, representing approximately 20% of all new breast cancer cases [1]. However, the risk for progression from DCIS to invasive cancer is still unclear, with estimates ranging from 14% to 53% [2]. In addition, up to 50% of lesions diagnosed as pure DCIS by core-needle biopsy will be upstaged to contain invasive disease at definitive surgery [3]. Methods that could predict an occult invasive component associated with this upstaging may affect treatment planning and avoid delays in definitive diagnosis.

Many studies have sought to identify preoperative predictors of DCIS upstaging. Various immunohistochemical biomarkers, histological features, and medical image findings have shown limited predictive power [3-13]. In contrast, computer extracted features, such as those designed for breast cancer screening [14-18], may be a promising alternative because of the quantitative and reproducible methodology. Previously, our group

[a]Carl E. Ravin Advanced Imaging Laboratories, Department of Radiology, Duke University School of Medicine, Durham, North Carolina.

[b]Department of Surgery, Duke University School of Medicine, Durham, North Carolina.

[c]Biodesign Center for Personalized Diagnostics and School of Life Sciences, Arizona State University, Tempe, Arizona.

[d]Centre for Evolution and Cancer, Institute of Cancer Research, London, United Kingdom.

Corresponding author and reprints: Bibo Shi, PhD, 2424 Erwin Road, Suite 302, Durham, NC 27705; e-mail: bibo.shi@duke.edu.

This work was supported in part by grant R01-CQA185138 from the National Cancer Institute and grant W81XWH-14-1-0473 from the US Department of Defense Breast Cancer Research Program. The authors have no conflicts of interest related to the material discussed in this article.

1546-1440/17/$36.00 ■ https://doi.org/10.1016/j.jacr.2017.11.036

[19] demonstrated that computer vision (CV) mammographic features can be used to predict DCIS upstaging with performance comparable with that of a radiologist. Nevertheless, this process of feature engineering is time consuming and might not capture all the image information.

Deep learning, especially deep convolutional neural networks (CNNs), has emerged as a promising approach for many image recognition or classification tasks [20], demonstrating human or even superhuman performance [21]. Typically deep learning requires training on very large image data sets with appropriate labeling and has been applied to several areas in medical imaging [22,23]. Deep CNN models learn a multiple-level, latent feature representation during the training procedure, without any input of prior expert knowledge. Used as a feature extractor, some pretrained CNN models can match or surpass the performance of domain-specific, handcrafted features [24-27]. Several recent studies applied this to medical tasks, including classification of images from chest radiography, chest CT, otoscopy, and endoscopy [28-32].

The purpose of this investigation was to determine whether DCIS upstaging can be predicted using deep features extracted from digital mammograms by a pretrained deep CNN. Additionally, we provide a head-to-head performance comparison between these nonmedical deep features and handcrafted CV features developed with breast cancer domain knowledge.

METHODS

Digital Mammogram Data Set

The study was approved by the institutional review board with a waiver of the requirement to obtain informed consent. We identified women aged 40 years or greater with stereotactic biopsy–proven DCIS presenting with only calcifications. Exclusion criteria included the presence of any masses, asymmetries, or architectural distortion; history of breast cancer or prior surgery; and presence of microinvasion at the time of initial biopsy. Ninety-nine subjects met these criteria, 25 of whom were upstaged. We collected the diagnostic digital magnification views, all produced by GE Senographe Essential systems (GE Healthcare, Little Chalfont, United Kingdom).

Extracting Handcrafted CV Features as Reference

As the reference baseline, we previously presented a model [19] based on handcrafted CV features. Three types of mammographic features were extracted from segmented individual microcalcifications (MCs) and the whole cluster for each DCIS lesion, including (1) shape features to describe the morphology and size of MCs and clusters, (2) topological features from weighted graphs associated with the clusters, and (3) texture features such as from gray-level co-occurrence matrices. Within a cluster, 25 CV features were computed for individual MCs. To describe the whole cluster, four global statistical measures were computed across all MCs: mean, standard deviation, minimum, and maximum. Overall, we obtained a set of 113 handcrafted CV features for each subject: 25×4 on the basis of individual MCs and 13 cluster features.

Extracting Deep Features Using a Pretrained CNN Model

In this study, we selected the widely used deep CNN model VGG [33] (specifically the configuration with 16 layers, VGG-16) because it has been very successful in many different localization and classification tasks [26,34-36]. Moreover, the VGG model adopts the most straightforward CNN architecture, which facilitates the extraction of deep features from multiple levels. As shown in Figure 1, the VGG-16 model consists of 16 weight layers, including 13 convolutional layers with filter size of 3×3, and 3 fully connected layers at the end. The convolutional layers are divided into five groups, and each group is followed by a max-pooling layer. The number of filters of convolutional layer group starts from 64 in the first group and then increases by a factor of 2 after each max-pooling layer, until it reaches 512. The overall trainable parameters are more than 130 million.

We used the VGG-16 model with parameters pretrained on the ImageNet data set [37], which is a large database with more than 20 million images arranged into more than 15,000 nonmedical concepts and categories (eg, animals, plants, instruments). During training, ImageNet images are usually resized into square, low-resolution ($224 \times 224 \times 3$ RGB) versions. Following that convention resulted in consistently poor performance for this study, however, because that resizing led to sacrificing the resolution or distorting the underlying topology of each lesion. Therefore, we directly input the region of interest (ROI) from the digital magnification view at its full resolution into the pretrained VGG-16 model. This means the input image size varies because of different lesion sizes. Accordingly, the last three fully connected layers cannot be used to extract features, and we only extracted deep features from convolutional layers,